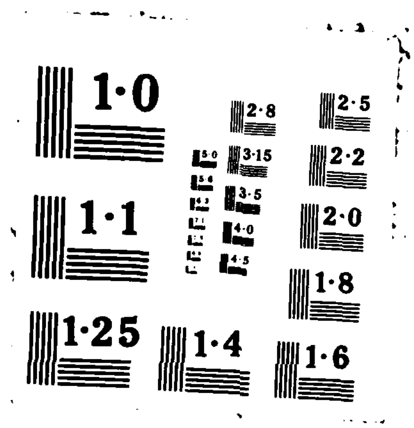
AD-A189 004 PROCEEDINGS OF THE WORKSHOP ON THE ASSESSMENT OF CREW WORKLOAD MEASUREMEN..(U) DOUGLAS AIRCRAFT CO LONG BEACH CA M A BIFERNO ET AL. JUN 87 AFWAL-TR-87-3043-VOL-1 1/4

UNCLASSIFIED F33615-86-C-3600 F/G 5/9 NL


1·0
2·8
2·5
3·15
2·2
3·5
1·1
2·0
4·0
4·5
1·8
1·25
1·4
1·6

(2)

DTIC FILE COPY

AFWAL-TR-87-3043, Vol I

PROCEEDINGS OF THE WORKSHOP ON THE ASSESSMENT OF CREW WORKLOAD MEASUREMENT METHODS, TECHNIQUES AND PROCEDURES

VOLUME I - Preliminary Selection of Measures

Dr M. A. Biferno
DOUGLAS AIRCRAFT COMPANY
3855 LAKEWOOD BLVD
LONG BEACH, CALIFORNIA 90846

George Boucek, Jr
BOEING COMMERCIAL AIRPLANE COMPANY
P. O. BOX 3707
SEATTLE, WASHINGTON 98124-2207

JUNE 1987

FINAL REPORT FOR PERIOD 24-25 FEBRUARY 1987

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

FLIGHT DYNAMICS LABORATORY
AIR FORCE WRIGHT AERONAUTICAL LABORATORIES
AIR FORCE SYSTEMS COMMAND
WRIGHT-PATTERSON AIR FORCE BASE, OHIO 45433-6553

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

AD-A189004

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION: Unclassified

1b. RESTRICTIVE MARKINGS:

2a. SECURITY CLASSIFICATION AUTHORITY:

2b. DECLASSIFICATION/DOWNGRADING SCHEDULE:

3. DISTRIBUTION/AVAILABILITY OF REPORT: Approved for public release; distribution unlimited.

4. PERFORMING ORGANIZATION REPORT NUMBER(S):

5. MONITORING ORGANIZATION REPORT NUMBER(S): AFWAL-TR-87-3043, Vol. I

6a. NAME OF PERFORMING ORGANIZATION: Douglas Aircraft Company

6b. OFFICE SYMBOL (If applicable):

7a. NAME OF MONITORING ORGANIZATION: Flight Dynamics Laboratory (AFWAL/FI6R)

6c. ADDRESS (City, State and ZIP Code):
3855 Lakewood Boulevard
Long Beach, California 90846

7b. ADDRESS (City, State and ZIP Code): WPAFB OH H5434-6553

8a. NAME OF FUNDING/SPONSORING ORGANIZATION:

8b. OFFICE SYMBOL (If applicable):

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER: F33615-86-C-3600

8c. ADDRESS (City, State and ZIP Code):

10. SOURCE OF FUNDING NOS.

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT NO.
62201F	2403	04	67

11. TITLE (Include Security Classification): (on back)

12. PERSONAL AUTHOR(S): Douglas Aircraft and Boeing Commercial Airplane

13a. TYPE OF REPORT: Final

13b. TIME COVERED: FROM 24Feb87 TO 25Feb87

14. DATE OF REPORT (Yr., Mo., Day): 87 June

15. PAGE COUNT: 373

16. SUPPLEMENTARY NOTATION: Supported by FAA and USAF

17. COSATI CODES

FIELD	GROUP	SUB. GR.
05	09	

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number):
Subjective measures
Performance measures
Physiological measures

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Report covers USAF/FAA review of workload measurement methods: validity, reliability, and applicability workshop.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT: UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☐ DTIC USERS ☐

21. ABSTRACT SECURITY CLASSIFICATION: Unclassified

22a. NAME OF RESPONSIBLE INDIVIDUAL: H.G. Britten-Austin

22b. TELEPHONE NUMBER (Include Area Code): (513)255-8259

22c. OFFICE SYMBOL: AFWAL/FIGR

DD FORM 1473, 83 APR   EDITION OF 1 JAN 73 IS OBSOLETE.   UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

(11) Title - Proceedings of the Workshop on the Assessment of Crew Workload measurement Methods, Techniques, and Procedures. Volume I. Preliminary selection of measures.

ACKNOWLEDGEMENTS

EDITED BY:

Biferno, M. A.	Douglas Aircraft Company Long Beach, California
Boucek, G. P.	Boeing Commercial Airplane Company Seattle, Washington
Hayward, K. B.	California State University Long Beach, California
Williams, K. N.	Douglas Aircraft Company Long Beach, California
Corwin, W. H.	Douglas Aircraft Company Long Beach, California
Sandry-Garza, D. L.	Boeing Commercial Airplane Company Seattle, Washington
Barbato, J.	California State University Long Beach, California
Logan, A.	Boeing Commercial Airplane Company Seattle, Washington
Dolan, M.	California State University Long Beach, California
D'Huyvetter, K.	California State University Long Beach, California

TABLE OF CONTENTS

USAF/FAA Review of Workload Measurement Methods:
Validity, Reliability and Applicability Workshop

February 24 and 25, 1987

Attendee List

Dr. Randall B. Aust
McDonnell Douglas Helicopter Co.
Building 531, c240
5000 East McDowell Road
Mesa, AZ 85205

Janet Barbato
Calif. State University, Long Beach
(Douglas Aircraft Co.)
Long Beach, CA 90802

Dr. R.M. Barnes
CAA House
Civil Aviation Authority
45-49 Kingsway
London, WC2B 6 TE England

Dr. Robert P. Bateman
Boeing Military Airplane Company
P.O. Box 7730, M/S 76-23
Wichita, KS 67277-7730

Mr. V. Battiste
NASA Ames Research Center
Attn: MS-239-3
Moffett Field, CA 94035

Dr. Michael A. Biferno
McDonnell Douglas Corporation
E21 MC 35-36
3855 Lakewood Boulevard
Long Beach, CA 90846

Dr. Richard Blomberg
Dunlap and Associates East Inc.
17 Washington Street
Norwalk, CT 06854

Michael R. Bortolussi
NASA Ames Research Center
MS 239-3
Moffett Field, CA 94035

George Boucek
10008 152nd E.
Payallup, WA 98373

Sqn Ldr Harry Britten-Austin
AFWAF/FIGR
Wright-Patterson Air Force Base
Dayton, OH 45433

Larry C. Butterbaugh
AFWAL/FIGR
Wright-Patterson Air Force Base
Dayton, OH 45433

Dr. Richard E. Christ
ARI Field Unit
P.O. Box 6057
Ft. Bliss, TX 79906

Dr. William Corwin
McDonnell Douglas Corporation
E21 MC 35-36
3588 Lakewood Boulevard
Long Beach, CA 90846

Dr. William L. Derrick
The Center for Creative Leadership
5000 Laurindra Dr. PO Box P-1
Greensboro, NC 27402

Dr. Michael Dresel
Lockheed Georgia Company
Dept. 72-09 Zone 419
Marietta, GA 30063

Dr. F. Thomas Eggemeier
Department of Psychology
University of Dayton
300 College Park
Dayton, OH 45469

Delmar M. Fadden
Boeing Commercial Airplane Co.
P.O. Box 3707, MS 77-70
Seattle, WA 98124-2207

Dr. Richard F. Gabriel
McDonnell Douglas Corporation
Dept E21 MC 35-36
3855 Lakewood Boulevard
Long Beach, CA [illegible]

Dr. Daniel Gopher
Technion-IIT
Faculty of Industrial Eng. & Mgmt.
Haifa 32000
Israel

Dr. Peter Hancock
University of Southern California
Dept. of Safety Sciences
Los Angeles, CA 90089-0021

Sandra G. Hart
NASA, Ames Research Center
MS 239-3
Moffett Field, CA 94035

Kathleen Hayward
Calif. State University, Long Beach
(Douglas Aircraft Co.)
Long Beach, CA 90802

Dr. Peter V. Hwoschinsky
FAA APM-430
800 Independence Avenue South West
Washington D.C. 20591

Dr. Richard S. Jensen
Ohio State University
Dept. of Aviation
Box 3022
Columbus, OH 43210

Dr. Arthur F. Kramer
University of Illinois
Psychology Department
603 East Daniel Street
Champaign, IL 61820

Aileen Logan
Boeing Commercial Airplane Co.
P.O. Box 3707, MS 66-25
Seattle, WA 98124-2207

George Lyddane
ANM-104L
4344 Donald Douglas Drive
Long Beach, [illegible]

Dr. Thomas M. McCloy
United States Air Force Academy
Behavioral Science & Leadership Dept.
Colorado Springs, CO 80840

Dr. Henry W. Mertens
Human Performance Research Unit
FAA/CAMI
Box 25082 AAC-118
Oklahoma City, OK 73125

Dr. Sam Metalis
McDonnell Douglas Corporation
E21 MC 35-36
3855 Lakewood Boulevard
Long Beach, CA 90846

Dr. Thomas R. Metzler
U.S. Army Research Institute
ARI AVSCOM PERI-IRA
4300 Goodfellow Boulevard
St. Louis, MO 63120-1798

Dr. Neville Moray
University of Toronto
Department of Industrial Engineering
Toronto, Ontario Canada M5S 1 A4

Dr. Robert D. O'Donnell
NTI Corporation
4130 Linden Dr., Suite 235
Claypool Building
Dayton, OH 45432

Charles Overbey
Science & Advanced Technology Office
Federal Aviation Administration
Washington D.C. 20591

Don L. Parks
Boeing Military Airplane Company
P.O. Box 3707, M/S 33-22
Seattle, WA 98124-2207

Dr. Alan Pope
NASA Langley Research Center
MS 152E
Hampton, VA 23665

Dr. Gary B. Reid
USAF/AMRL
Wright-Patterson Air Force Base
Dayton, OH 45433

John M. Reising
AFWAL/FIGR
Wright-Patterson Air Force Base
Dayton, OH 45433-6553

Dr. Alan Roscoe
Britannia Airways Ltd.
Luton Airport
Bedfordshire LU2 9ND United Kingdom

John Rye
Civil Aviation Authority
CAA House
45-59 Kingsway
London WC2B 6TE United Kingdom

Diane L. Sandry-Garza
Boeing Commercial Airplane Co.
P.O. Box 3707, MS 66-25
Seattle, WA 98124-2207

Dr. Richard Schiffler
ASD ENEC
Wright-Patterson Air Force Base
Dayton, OH 45433

Dr. Samuel G. Schiflett
USAF SAM/VNCPL
Crew Technology Division
SAM/UNB Brooks AFB, TX 78235

Dr. Clark A. Shingledecker
NTI, Inc.
Suite 235
4130 Linden Ave.
Dayton, OH 45432

Dr. Kim Siler
Lockheed-California Co.
P.O. Box 551
Burbank, CA 91520

Dr. Don Smith
7310 Valburn Drive
Austin, TX 78731

Jean-Jacques Speyer
Airbus Industries
Performance & Operations Engin.
Flight Division HQ BP No33
31707 Blagnac Cedex, France

Dr. John A. Stern
Professor of Psychology
Washington University Box 1125
St. Louis, MO 63130

Guice Tinsley
FAA AFS 210
800 Independence Avenue S.W.
Washington D.C. 20591

Mr. W.A. Wainwright
BAE Civil Division
Hatfield
Hertfordshire AL10 9TL England

Dr. Larry Walrath
McDonnell Douglas Co.
Building 107 MS 10
P.O. Box 516
St. Louis, MO 63166

Dr. Christopher E. Wickens
University of Illinois
Department of Psychology
603 East Daniel Street
Champaign, IL 61820

Dr. Ken Williams
McDonnell Douglas Corporation
E21 MC 35-36
3588 Lakewood Boulevard
Long Beach, CA 90846

Dr. Glenn F. Wilson
USAF
AMRL, AAMRL/HEG
Wright-Patterson AFB, OH 45433-6573

USAF/FAA Review of Workload Measurement Methods:
Validity, Reliability, and Applicability

PROGRAM - WORKSHOP 1
February 24 and 25, 1987

DAY 1

TIME	EVENT
BALLROOM "A"	
8:00	Welcome from Douglas Aircraft Engineering. (Peterson)
8:10	Welcome from USAF/FAA sponsors. (Britten-Austin and Hwoschinsky)
8:20	Objectives of workshop. (Biferno)
8:30	What is workload certification? (Gabriel)
9:00	Workload Assessment and certification. (Fadden)
9:30	Methodological issues concerning workload measurement during certification. (Biferno, Boucek)
10:15	Break
10:30	Subjective workload measurement panel: A review of the evidence regarding validity and reliability. Reviewer Hart Reviewer Reid Discussant Gopher
12:00	Lunch - LOUNGE
1:00	Performance workload measurement panel: A review of the evidence regarding validity and reliability. Reviewer Wickens Reviewer Eggemeier Discussant McCloy
2:30	Break
2:45	Physiological workload measurement panel: A review of the evidence regarding validity and reliability. Reviewer Kramer Reviewer Wilson Discussant Stern

DAY 1 (continued)

TIME	EVENT
4:15	Review of criteria for FACT MATRICIES categorization. (Williams)
4:30	Review of participant Fact Matrices by workload panel. Three subgroups will be created with panel members as leaders. Subjective subgroup in room # 106 Performance subgroup in room # 136 (Wednesday in room # 121) Physiological subgroup in room #151
5:30	Adjourn.
7:00	Banquet - MONTEGO BAY ROOM

DAY 2

TIME	EVENT
8:00	Continue review of Fact Matrices by workload panels in subgroups. Cite evidence for additions or deletions to original matrix.
12:00	Lunch - LOUNGE
BALLROOM "A"	
1:30	Review Fact Matrices of Subjective Measures.
2:00	Review Fact Matricies of Performance Measures.
2:30	Review Fact Matricies of Physiological Measures.
3:00	Break
3:15	Briefing on how measures will be implemented in simulation scenario. (Corwin, Sandry-Garza)
4:15	Concluding remarks. (Boucek, Biferno)
4:30	Survey of attendees regarding the best workload measures.
4:45	Turn in survey and final version of Fact Matrices.
	Cash bar - MONTEGO BAY ROOM

OBJECTIVES

OBJECTIVE OF THE WORKSHOP

GATHER INFORMATION FROM WORKLOAD EXPERTS REGARDING WHICH MEASURES HAVE EVIDENCE SUPPORTING THEIR RELIABILITY OR VALIDITY.

OBJECTIVES OF PANEL DISCUSSIONS

PROVIDE AN INDEPENDENT REVIEW OF THE FACTS CONCERNING THE VALIDITY AND RELIABILITY OF WORKLOAD MEASURES.

OBJECTIVE OF THE SUBGROUP SESSIONS

PROVIDE A MEANS FOR SYSTEMATICALLY REVIEWING AND MODIFYING THE FACT MATRICES.

USAF/FAA Review of Workload Measurement Methods:
Validity, Reliability and Applicability Workshop

SUBGROUP PARTICIPATION LIST

SUBJECTIVE
(In room #106)

Dr. Randall Aust
Dr. R.M. Barnes
Dr. Robert P. Bateman
Mr. V. Battiste
Dr. M.A. Biferno
Dr. Richard Blomberg
Dr. Larry C. Butterbaugh
Dr. Richard E. Christ
Dr. Daniel Gopher (moderator-chair)
Ms. Sandra G. Hart (moderator)
Ms. Aileen Logan (assistant)
Mr. Thomas R. Metzler
Dr. Gary B. Reid (modearator)
Mr. John Rye
Ms. Diane L. Sandry-Garza

PERFORMANCE
(Tuesday in room #136)
(Wednesday in room #121)

Ms. Janet Barbato (assistant)
Mr. Michael R. Bortolussi
Dr. George Boucek
Dr. William Corwin
Dr. William Derrick
Dr. Michael Dresel
Dr. F. Thomas Eggemeier (moderator)
Dr. Richard S. Jensen
Dr. Thomas McCloy (moderator-chair)
Mr. Don L. Parks
Mr. John M. Reising
Dr. Clark A. Shingledecker
Dr. Kim Siler
Mr. W.A. Wainwright
Dr. Christopher Wickens (moderator)

PHYSIOLOGICAL
(In room #151)

Dr. Peter Hancock
Ms. Kathleen Hayward (assistant)
Dr. Arthur F. Kramer (moderator)
Dr. Sam Metalis
Dr. Neville Moray
Dr. Robert D. O'Donnell
Dr. Alan Pope
Dr. Alan Roscoe
Dr. Richard Schiffler
Dr. Samuel G. Schiflett
Mr. Jean-Jacques Speyer
Dr. John A. STern (moderator-chair)
Dr. Larry Walrath
Dr. Ken Williams
Dr. Glenn F. Wilson (moderator)

OBSERVERS

Sqn Ldr Harry Britten-Austin
Mr. Delmar M. Fadden
Dr. Richard F. Gabriel
Dr. Peter V. Hwoschinsky
Mr. George Lyddane
Dr. Henry W. Mertens
Mr. Charles Overbey
Mr. Don Smith
Mr. Guice Tinsley

WORKLOAD FACT MATRIX REVIEW
PANEL AGENDA
(suggested)

TIME	ITEM	ITEM DESCRIPTION
1630-1730	1.	CRITERIA FOR FACT MATRIX REVIEW a. Reliability b. Validity c. Empirical data d. Capability for flight use e. Rejection/Addition rationale and support
0800-1030	2.	REVIEW OF EACH MEASURE ON FACT MATRIX Process: 1) Pick workload measure for discussion 2) Identify empirical material which provide evidence on validity and reliability of measure 3) Solicit change recommendations 4) Review evidence for change recommendations 5) Speaker Panel form recommendation 6) Repeat process for next measure
1030-1130	3.	REVIEW MEASURES NOT PREVIOUSLY INCLUDED IN FACT MATRICES Process: 1) Identify a new measure for inclusion 2) Review evidence for addition of measure 3) Speaker Panel form recommendation 4) Repeat process until no additional measures identified
1130-1200	4.	FORMULATE SUMMARY REPORT Summarize: additional reference items delete reference items additional measures

WORKLOAD FACT MATRIX REVIEW

The purpose of the Subgroup reviews is to modify the FACT MATRICES in a systematic and orderly fashion. The FACT MATRICES are basically locators for Reliability and Validity information for each Workload measurement type. We want you to add reference work to the FACT MATRICES if they contain Reliability and Validity information. We are not asking your panel to judge the quality of the information, but determine whether the material addresses the measure's Reliability or Validity. The evidence can be weak and still be acceptable for inclusion in the FACT MATRIX. On the other hand, if we entered references into the FACT MATRICES which are not appropriate or correct, we ask you to delete those items. Regardless of the modifications recommended by your subgroup, evidence must be given for the addition or deletion of items. It is up to the Speaker Panel to decide on the acceptability of the evidence and on the decision to add or subtract from the subgroup sessions and direct the decision process that the group employs to modify the FACT MATRICES.

We ask that each subgroup first employ the Anastasi and Guilford definitions of Validity and Reliability to the workload literature. We are aware that the workload field has not generally focused its resources on demonstrations of reliability because there is disagreement on the definitions and content areas of the workload construct. Therefore, after addressing the literature using the Anastasi and and Guilford criteria, you are free to employ the validity and reliability definitions of your choice to support the contention that a particular measure is valid and reliable. We only ask that you explicitly define these definitions in your justification for that measure. Remember the studies suitable for supporting validity or reliability must be empirical, not review or theory articles.

VALIDITY DEFINITIONS

1. FACE VALIDITY: "... pertains to whether the test 'loods valid' to the subjects who take it, the administrative personnel who decide on its use, and other technically untrained observers."
IMPORTANCE: "Certainly if a test appears irrelevant, inappropriate, silly, or childish, the result will be poor cooperation, regardless of the actual validity of the test." ... "For example, it a test of simple arithmetic reasoning is constructed for use with machinists, the items should be worded in terms of machine operations rather than in terms of 'how many oranges can be purchased for 36 cents' of other traditional schoolbook problems."
REFERENCE: Anastasi, A. (1968). Psychological Testing. 3rd Edition, Macmillan, London, p. 104.

2. CONTENT VALIDITY: " ... involves essentially the systematic examination of the test content to determine whether it covers a representative sample of the behavior domain to be measured." ... "The content area to be tested must be systematically analysed to make certain that all major aspects are adequately covered by the test items, and in the correct proportions."
IMPORTANCE: " ... content validity depends on the relevance of the individual's test responses to the behavior area under consideration, rather than on the apparent relevance of item content."
REFERENCE: Anastasi, A. (1968). Psychological Testing. 3rd Edition, Macmillan, London, p. 100.

3. CONSTRUCT VALIDITY: " ... is the extent to which the test may be said to measure a theoretical construct or trait." ... "requires the gradual accumulation of information from a variety of sources. Any data throwing light on the nature of the trait under consideration and the conditions affecting its development and manifestations are grist for this validity mill."
IMPORTANCE: " ... construct validity is a comprehensive concept, which includes the other types. All specific techniques for establishing content and criterion-related validity ... could be listed again under construct validity."
REFERENCE: Anastasi, A. (1968). Psychological Testing. 3rd Edition, Macmillan, London, p. 114-5, 121.

4. PREDICTIVE VALIDITY: (CRITERION RELATED VALIDITY) " ... indicates the effectiveness of a test in predicting an individual's behavior in specific situations."
IMPORTANCE: "For this purpose, performance on the test is checked against a criterion, i.e., a direct and independent measure of that which the test is designed to to predict."
REFERENCE: Anastasi, A. (1968). Psychological Testing. 3rd Edition, Macmillan, London, p. 105.

RELIABILITY DEFINITIONS

1. TEST-RETEST RELIABILITY: " The key concept for this (test-retest) procedure is that of stability. It answers the question concerning how stable or dependable are the measurements over a period of time."

IMPORTANCE: "High reliability of this kind tells us that the individuals remain rather uniform, or maintain their rank positions in spite of changes, in whatever psychological functions this test measures."

REFERENCE: Guilford, J. P. (1954). Psychometric Methods, 2nd Edition, McGraw Hill, New York, p. 374.

2. SPLIT HALF RELIABILITY: "The information sought (from fractionation of a test into two or more parts) concerns the equivalence of parts for measurement purposes, or the internal consistency of the test."

IMPORTANCE: "In the case of the split-half method, the Spearman-Brown formula has usually been applied to estimate the reliability of the test of full length from the obtained estimate of correlation of a test of half length."

REFERENCE: Guilford, J. P. (1954). Psychometric Methods, 2nd Edition, McGraw Hill, New York, p. 373.

3. ALTERNATE FORMS RELIABILITY: "... method bears resemblances to both the internal consistency approach and the retest approach. the end result is an index of how equivelant the psychological-measurement content of one form of the test is with the content of another."

IMPORTANCE: "... the alternate-forms method indicates both the equivelance of content and stability of performance." ... " Some investigators prefer the alternate-forms type to the internal consistency type of coeffecient for the reason that they are interested in how much stability to expect of scores over time."

REFERENCE: Guilford, J. P. (1954). Psychometric Methods, 2nd Edition, McGraw Hill, New York, p. 374,5.

4. INTER-RATER RELIABILITY: " ... rater intercorrelations, which indicate the internal consistency among raters. Such correlations have usually been regarded as indices of rating reliability ..."

IMPORTANCE: If raters agree, demonstrate high intercorrelations, then the number of raters required to generate a significant result cna be reduced.

REFERENCE Guilford, J. P. (1954). Psychometric Methods, 2nd Edition, McGraw Hill, New York, p. 286,7.

ASSESSMENT FORM

Lib. No.______

1st Init.______

2nd Init.______

CHECK

A. Content (If none are checked QUIT)

___ 1 Workload Measurement Primary.

___ 2 Workload Measurement Secondary.

NUMBER OR NAME

B. Quality of Review

___ 1= Formal Review 2= Informal Review 3= No Reviewer

C. Quality of Data (If not 1-3 QUIT)

___ 1= Experiment(s) 2= Case Study(s) 3= Theory/Review
(Skip to F.) (Skip to H.)

D. ______________ Experiment Name if more than one (One per form)

E. Highest Fidelity of Experiment

___ 1= Actual Flight 2= Simulator 3= Applied lab. 4= Basic lab.

CHECK

F. Validities (If none are checked skip to G.)

___ 1 Content.

___ 2 Construct.

___ 3 Predictive.

MEMO

Independent
Variables:

G. Reliabilities

___ 1 Test-retest.

___ 2 Split half.

___ 3 Alternate forms.

___ 4 Inter rater.

H. Measure Types

Dependent
Variables:

Subjective.

___ 1 NASA Bipolar Scale

___ 2 SWAT

___ 3 WCI/TE

___ 4 Modified Cooper Harper

___ 5 Interviews

___ 6 Surveys

___ 7 Other Subjective Measures

Physiological.

___ 8 Body Fluid

___ 9 Brain Activity

___ 10 Heart

___ 11 Lung

___ 12 Muscle

___ 13 Skin

___ 14 Vision

___ 15 Voice

___ 16 Other Physiological Measures

Performance.

___ 17 Primary Task

    ___ 17a Time

    ___ 17b Position

    ___ 17c Event

___ 18 Normal Secondary Task

    ___ 18a Time

    ___ 18b Position

    ___ 18c Event

___ 19 Artificial Secondary Task

    ___ 19a Time

    ___ 19b Position

    ___ 19c Event

I. Workload Types
___ 4a Degree of Mental.
___ 4b Duration of Mental.
___ 4c Degree of Physical.
___ 4d Duration of Physical.

J. Workoad Functions
___ 1 Flight path control.
___ 2 Collision avoidance.
___ 3 Navigation.
___ 4 Communications.
___ 5 Operation and Monitoring.
___ 6 Command decisions.

K. Workload Factors (Task Demands)
___ 1 Normal Control.
___ 2 Normal Display.
___ 3 Normal Procedure.
___ 5 Normal Monitoring.
___ 8a Normal Communication.
___ 8b Normal Navigation.
___ 6 Non-normal Crew unavailability.
___ 7 Non-normal Automation.
___ 9 Non-normal Procedure.
___ 10 Non-normal Crew incapacitation.

FAR-25 WORKLOAD FACTOR 4c: DEGREE OF PHYSICAL

MEASURE	VALIDITY			RELIABILITY				APPLICABILITY
	CONTENT	CONSTRUCT	PREDICTIVE	TEST RETEST	SPLIT HALF	ALTERNATE FORMS	INTER RATER	
NASA Bipolar Scale	28 95 268 530	28 268 539				95		
SWAT	28 41 339 340 773	28 41 339 340 766 773	340 766 773			340 766		
WCI/TE	13 16 64 243 367 538	13 16 64 243 367	16 538	64	64	64	64	
Modified Cooper Harper	13 15 16 65 130(1) 159 175 183 225 243 344 367 436 459 492 496 509 538 540 583 669 735 762 778 793 799	13 15 16 65 130(1) 175 183 187 225 243 283 367 436 459 492 496 509 583 735 761 762 778 793	15 16 65 175 183 225 283 436 459 492 496 538 583 761 762 778	225 583		183 283 344 459 496 583 762 793	183 225 762	

USAF/FAA REVIEW OF WORKLOAD MEASUREMENT METHODS: VALIDITY, RELIABILITY, AND APPLICABILITY

FEBRUARY 24 AND 25, 1987

LONG BEACH, CALIFORNIA

CC1335 82

OBJECTIVES

OBJECTIVE OF THE WORKSHOP

- GATHER INFORMATION FROM WORKLOAD EXPERTS REGARDING WHICH MEASURES HAVE EVIDENCE SUPPORTING THEIR RELIABILITY OR VALIDITY

OBJECTIVE OF PANEL DISCUSSIONS

- PROVIDE AN INDEPENDENT REVIEW OF THE FACTS CONCERNIG THE VALIDITY AND RELIABILITY OF WORKLOAD MEASURES

OBJECTIVE OF THE SUBGROUP SESSIONS

- PROVIDE A MEANS FOR SYSTEMATICALLY REVIEWING AND MODIFYING THE FACT MATRICES

CC1336 01

PROGRAM SCOPE

IDENTIFY EXISTING WORKLOAD MEASURES

EVALUATE MOST PROMISING MEASURES

EVALUATE MEASURES SUITABLE FOR FULL-MISSION SIMULATION OR IN-FLIGHT

EVALUATE TRANSPORT ENVIRONMENT

WORKSHOP OUTPUT

COPIES OF VIEWGRAPH PRESENTATIONS

FINAL FORM OF FACT MATRICES

SURVEY OF WORKSHOP PARTICIPANTS

CC1336 83

SOME
CONSIDERATIONS REGARDING CREW WORKLOAD EVALUATION DURING AIRCRAFT CERTIFICATION

CA1458.01

A DEFINITION

CERTIFICATION IS THE FORMAL PROCESS THROUGH WHICH THE AIRFRAME MANUFACTURER AND THE FAA ENSURE COMPLIANCE WITH FEDERAL AIR REGULATIONS

CA1458.04

CODE OF FEDERAL REGULATIONS

(GENERAL AND PERMANENT RULES OF THE U.S. GOVERNMENT EXECUTIVE DEPARTMENTS/AGENCIES)

• 50 TITLES	TITLE 14 - AERONAUTICS AND SPACE
• SUBDIVIDED INTO CHAPTERS	CHAPTER I - THE FEDERAL AVIATION ADMINISTRATION
SUBCHAPTERS	SUBCHAPTER C - AIRCRAFT
PARTS	PART 25 - AIRWORTHINESS STANDARDS-TRANSPORT CATEGORY AIRCRAFT
SUBPARTS	SUBPART G - OPERATING LIMITATIONS AND INFORMATION
PARAGRAPHS	PARAGRAPH 25.1523 - MINIMUM FLIGHT CREW
APPENDICES AND AMENDMENTS	APPENDIX D - CRITERIA FOR DETERMINING MINIMUM FLIGHT CREW

CA1458.87

WHO/WHAT GETS CERTIFIED

AIRCRAFT (TRANSPORT)	PART 25
AIRCRAFT NOISE STANDARDS	PART 36
PILOTS AND FLIGHT INSTRUCTORS	PART 61
AIRLINE OPERATORS	PART 121
PILOT TRAINING CENTERS	PART 141
AIRCRAFT REPAIR STATIONS	PART 145

CA1458 06

PARTICIPANTS IN THE CERTIFICATION PROCESS

WHO CERTIFIES?

THE FAA:

- ENGINEERS
- PILOTS
- NATIONAL RESOURCE SPECIALISTS
- DERs
- POLICY AND PROCEDURE SPECIALISTS
- ADMINISTRATORS
- MAINTENANCE SPECIALISTS

WHO COMPLIES?

- AIRFRAME MANUFACTURERS
- EQUIPMENT MANUFACTURERS
- AIRLINES

WHO CRITIQUES?

- EVERYONE, AND LOUDLY

CA1458.05

REQUIREMENTS FOR CERTIFICATION

DEMONSTRATE TO THE FAA'S SATISFACTION THAT THE DESIGN IS IN COMPLIANCE WITH APPLICABLE

- FARs
- SPECIAL CONDITIONS
- OPTIONAL CONDITIONS

CA1458.08

METHODS OF DEMONSTRATING COMPLIANCE

MEET ADVISORY CIRCULAR STANDARDS	ISSUED BY FAA. NOT LEGALLY REQUIRED, BUT. . . .
ANALYSIS DEMONSTRATION TEST	MANUFACTURER PREPARES PLAN WHICH IS REVIEWED AND REVISED UNTIL APPROVED BY FAA. SUBMITTALS SUBSTANTIATE THA THE APPROVED PLAN HAS BEEN FOLLOWED

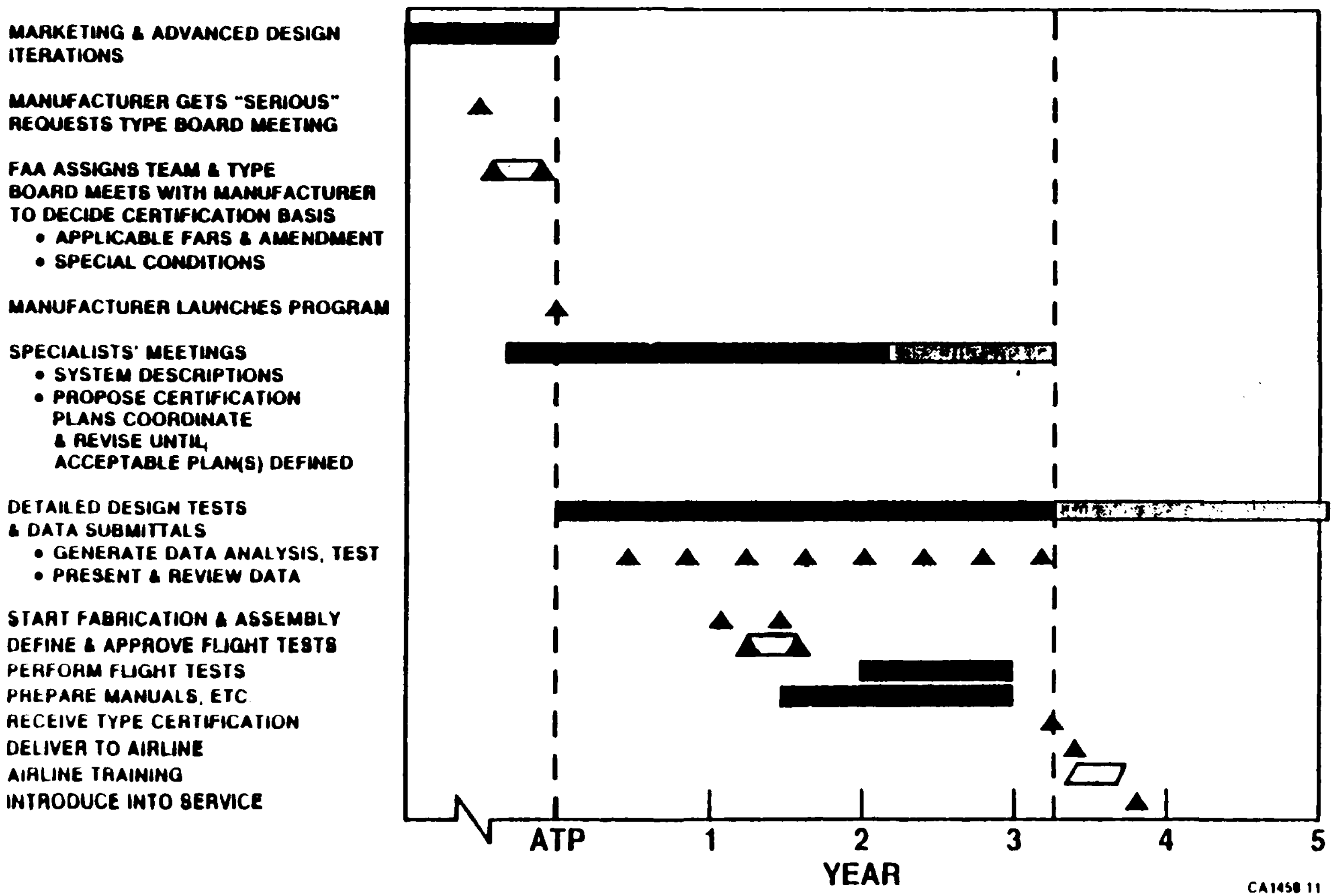
CA1458.09

REPRESENTATIVE FAR PARAGRAPHS

METHOD	EXAMPLE
DEMONSTRATION (1) (SUBJECTIVE ASSESSMENT)	§25.771. PILOT COMPARTMENT. MUST ALLOW THE MINIMUM FLIGHT CREW TO PERFORM THEIR DUTIES WITHOUT UNREASONABLE CONCENTRATION OR FATIGUE.
DEMONSTRATION (2)	§25.803. EMERGENCY EVACUATION. IT MUST BE SHOWN BY ACTUAL DEMONSTRATION THAT THE MAXIMUM SEATING CAPACITY, INCLUDING CREW MEMBERS, CAN BE EVACUATED FROM THE AIRPLANE TO THE GROUND WITHIN 90 SECONDS.
ANALYSIS (FORMAL APPLICATION OF KNOWN SCIENTIFIC LAWS TO REACH CONCLUSION)	§25.581. LIGHTNING PROTECTION. THE AIRPLANE MUST BE PROTECTED AGAINST CATASTROPHIC EFFECTS OF LIGHTNING FOR METALLIC COMPONENTS. COMPLIANCE MAY BE SHOWN BY (1) BONDING THE COMPONENTS PROPERLY TO THE AIRFRAME, OR (2) DESIGNING THE COMPONENTS SO THAT A STRIKE WILL NOT ENDANGER THE AIRCRAFT.
TEST (FLIGHT)	§25.203. STALL CHARACTERISTICS. FOR LEVEL WING STALLS, THE ROLL OCCURRING BETWEEN THE STALL AND THE COMPLETION OF THE RECOVERY MAY NOT EXCEED APPROXIMATELY 20 DEGREES.
TEST	§25.651. PROOF OF STRENGTH. LIMIT LOAD TESTS OF CONTROL SURFACES ARE REQUIRED. THESE TESTS MUST INCLUDE THE HORN OR FITTING TO WHICH THE CONTROL SYSTEM IS ATTACHED.

CA1458.10

A REPRESENTATIVE PROGRAM PROGRESSION



WHAT CERTIFICATION IS NOT

TRIVIAL, EASY, OR INEXPENSIVE

THE ONLY CHECK-AND-BALANCE ON SAFETY
- MANUFACTURER'S EMPHASIS
- DOES KEEP FROM BEING OVERCONFIDENT

A GUARANTEE OF PERFECT SAFETY

TOTALLY FREE OF SUBJECTIVITY
- 2.5 g WINS
- 1.5 g SAFETY FACTOR
- HANDLING QUALITIES
- 90-SECOND EVACUATION

PERFECTLY CONSISTENT
- WITHIN PARAGRAPHS
- BETWEEN PROGRAMS

CA1458.12

CREW WORKLOAD IS A GROWING CONSIDERATION IN CERTIFICATION

CHANGING TECHNOLOGY
- AUTOMATION

CHANGING ENVIRONMENT
- ATC
- WEATHER

ECONOMIC CONSIDERATIONS

AIRLINE PILOT ASSOCIATION POSITION

PRESIDENTIAL COMMISSION RECOMMENDATIONS

MILITARY SYSTEMS

INCREASING AWARENESS OF HUMAN FACTORS ISSUES

CA1458.03

WORKLOAD ASSESSMENT - WHY CURRENT APPROACH IS NOT ADEQUATE

HEAVY RELIANCE ON SUBJECTIVE ASSESSMENT

- PILOTS' JUDGEMENTS VARY
- PROBLEMS MAY BE DISCOVERED LATE
- FAA PILOTS HAVE HEAVY RESPONSIBILITY

TASK/TIMELINE ANALYSIS HAS LIMITATIONS

- SERIAL/PARALLEL ACTIVITIES
- CONTINUOUS ACTIVITIES
- MENTAL WORKLOAD

LACK OF STANDARDIZATION

LACK OF EXPLICIT STANDARDS

CA1458.15

GOALS OF THIS PROGRAM

LONG TERM

- DEVELOP VALID, RELIABLE SENSITIVE MEASURES OF WORKLOAD THAT CAN BE READILY USED TO FACILITATE DESIGN AND CERTIFICATION OF COMPLEX SYSTEMS
- DEVELOP WORKLOAD STANDARDS THAT PROVIDE PASS-FAIL CRITERIA BASED ON DEMONSTRATED EFFECTS OF WORKLOAD ON PERFORMANCE

NEAR TERM

- PROVIDE CLEAR, CONCRETE, CONCISE GUIDELINES THAT WILL RESULT IN THE APPLICATION OF THE BEST APPROACH TO WORKLOAD ASSESSMENT BASED ON AVAILABLE KNOWLEDGE

CA1458 02

FAR RELEVANT TO CREW WORKLOAD

δ 25.1523 MINIMUM FLIGHT CREW

THE MINIMUM FLIGHT CREW MUST BE ESTABLISHED SO THAT IT IS SUFFICIENT FOR SAFE OPERATION, CONSIDERING:

(a) THE WORKLOAD ON INDIVIDUAL CREW MEMBERS;

(b) THE ACCESSIBILITY AND EASE OF OPERATION OF NECESSARY CONTROLS BY THE APPROPRIATE CREW MEMBER; AND

(c) THE KIND OF OPERATION AUTHORIZED UNDER δ 25.1525. THE CRITERIA USED IN MAKING THE DETERMINATIONS REQUIRED BY THIS SECTION ARE SET FORTH IN APPENDIX D

CA1450.13

APPENDIX D - SUMMARY

CRITERIA FOR DETERMINING MINIMUM FLIGHT CREW

(a) BASIC WORKLOAD FUNCTIONS
- (1) FLIGHT PATH CONTROL
- (2) COLLISION AVOIDANCE
- (3) NAVIGATION
- (4) COMMUNICATIONS
- (5) OPERATING AND MONITORING OF AIRCRAFT ENGINES AND SYSTEMS
- (6) COMMAND SYSTEMS

(b) WORKLOAD FACTORS
- (1) ACCESSIBILITY, EASE, AND SIMPLICITY OF OPERATION . . . CONTROLS
- (2) *ACCESSIBILITY AND CONSPICUITY OF* . . . INSTRUMENTS . . .
- (3) NUMBER, URGENCY, AND COMPLEXITY OF OPERATING PROCEDURES
- (4) THE DEGREE AND DURATION OF CONCENTRATED MENTAL AND PHYSICAL EFFORT IN NORMAL OPERATION IN DIAGNOSING AND COPING WITH MALFUNCTIONS AND EMERGENCIES
- (5) THE EXTENT OF REQUIRED MONITORING OF . . . SYSTEMS
- (6) THE ACTIONS REQUIRING A CREW MEMBER TO BE UNAVAILABLE AT HIS ASSIGNED STATION . . .
- (7) THE DEGREE OF AUTOMATION PROVIDED . . .
- (8) THE COMMUNICATIONS AND NAVIGATIONS WORKLOAD
- (9) THE POSSIBILITY OF INCREASED WORKLOAD ASSOCIATED WITH AN EMERGENCY LEADING TO OTHER EMERGENCIES
- (10) INCAPACITATION OF A CREW MEMBER

(c) KIND OF OPERATION AUTHORIZED (ASSUMED IFR, UNLESS . . .)

CA1458.14

DESIRABLE ATTRIBUTES OF WORKLOAD MEASURE

VALIDITY - WILL THE MEASURE REALLY REFLECT "WORKLOAD"?

RELIABILITY - WILL THE MEASURE BE CONSISTENT?

SENSITIVITY - WILL SMALL BUT MEANINGFUL DIFFERENCES BE DISCRIMINATED?

UTILITY - CAN MEASUREMENT PROCESS BE PERFORMED IN REPRESENTATIVE ENVIRONMENTS?

COMPREHENSIVENESS - CAN DIFFERENT WORKLOAD TYPES BE COMPARED?

ACCEPTANCE - WILL ALL IMPACTED PARTIES ACCEPT THE RESULTS?

REASONABLE COSTS

QUICK TURNAROUND

CAN BE APPLIED THROUGHOUT DESIGN PROCESS

PROVIDES DIAGNOSTIC CAPABILITY (HELPS TO DETERMINE CHANGES TO CORRECT CONDITION

NON-INTERFERING/UNOBTRUSIVE

ACCOMMODATES INDIVIDUAL DIFFERENCES

CA1458 18

NEED FOR WORKLOAD STANDARD(S)

CERTIFICATION IMPLIES A PASS-FAIL CRITERION

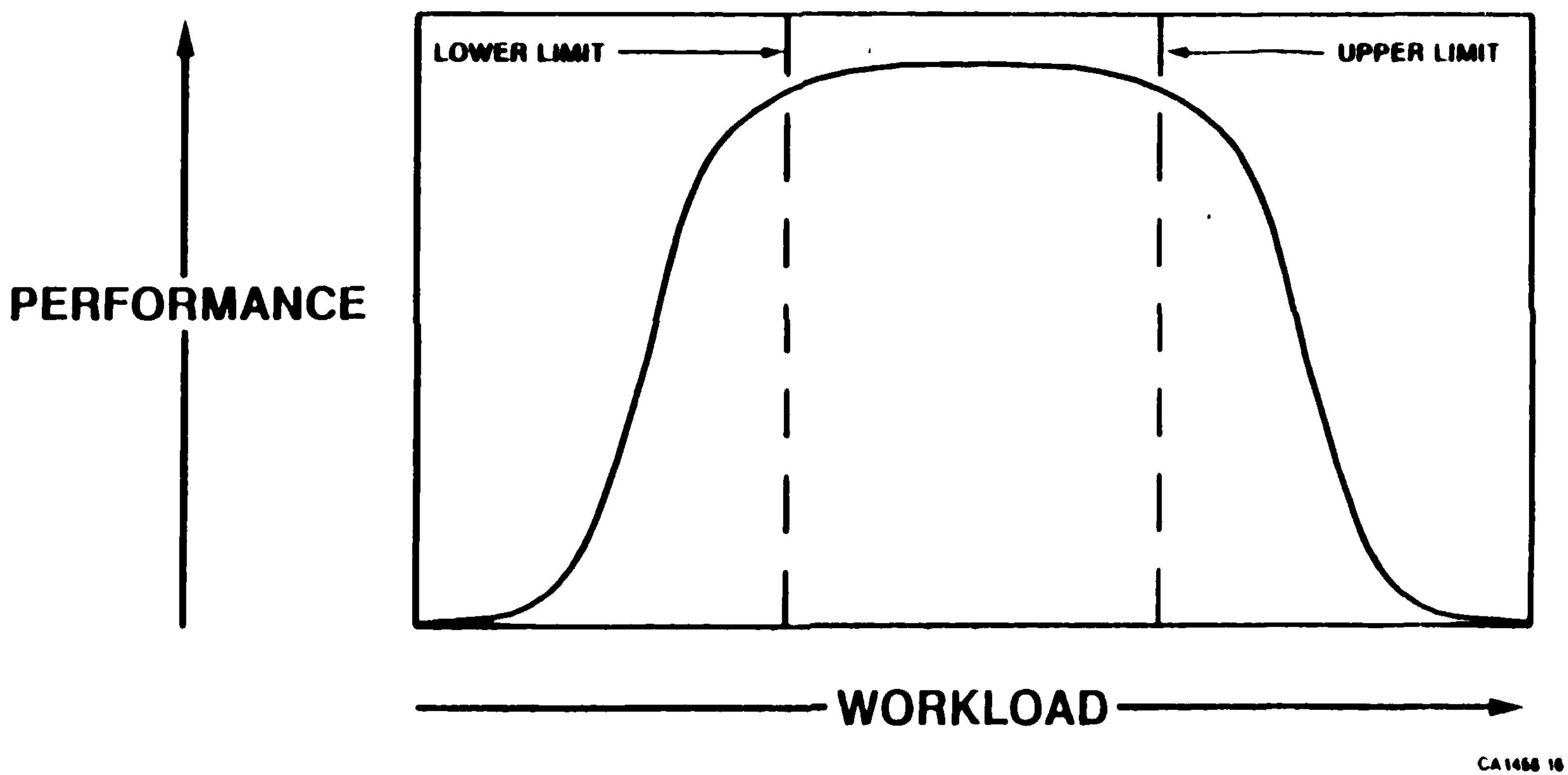
ACCURATE, RELIABLE MEASUREMENT IS FIRST STEP

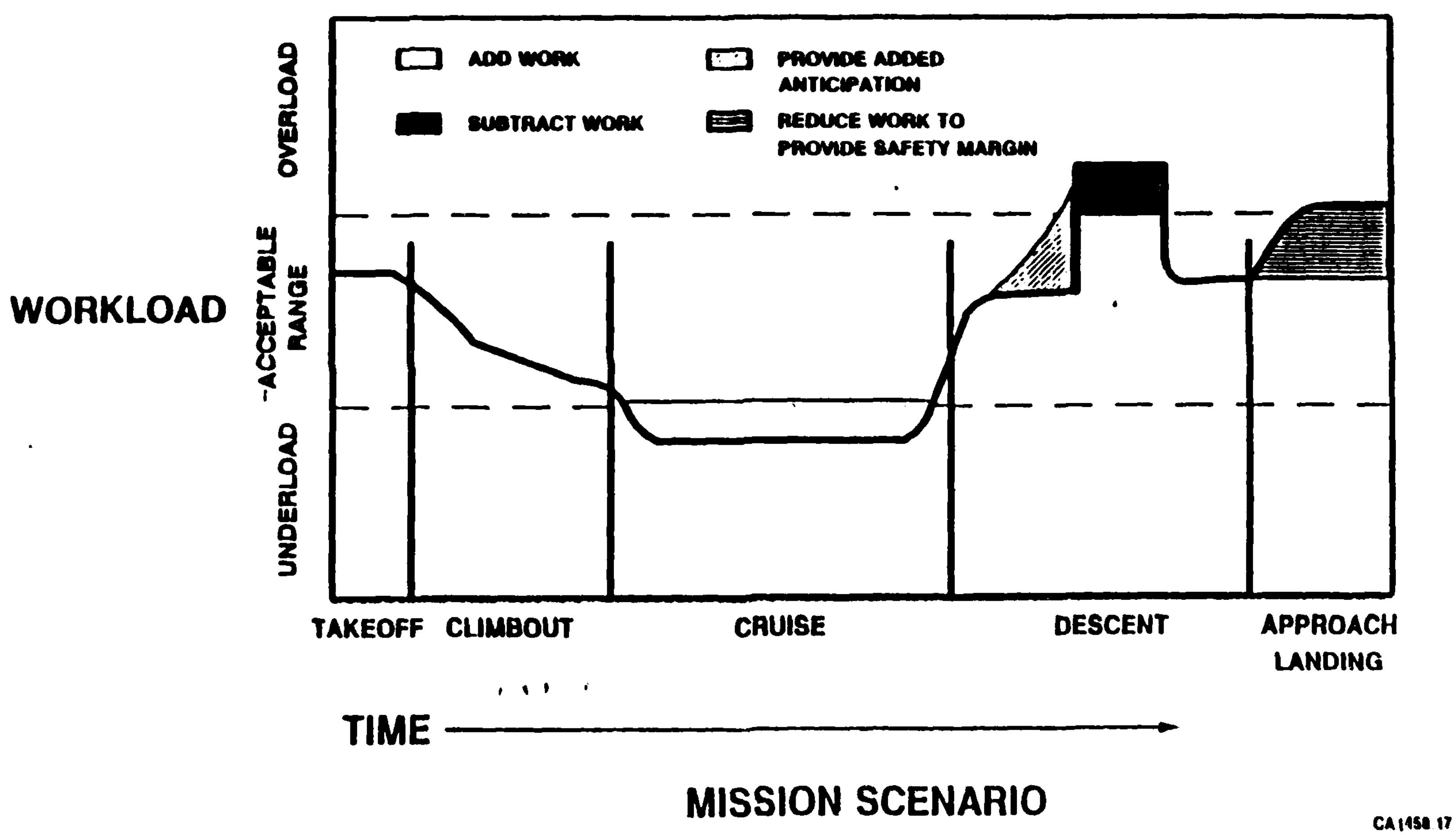
MUST DETERMINE WHEN WORKLOAD IS OUTSIDE ACCEPTABLE LIMITS

FOR CERTIFICATION, "ACCEPTABLE" RELATES TO SAFETY OF FLIGHT

CA1458 19

HYPOTHESIZED RELATIONSHIP BETWEEN WORKLOAD AND PERFORMANCE




ADD WORK
SUBTRACT WORK
PROVIDE ADDED ANTICIPATION
REDUCE WORK TO PROVIDE SAFETY MARGIN
WORKLOAD
OVERLOAD
ACCEPTABLE RANGE
UNDERLOAD
TAKEOFF
CLIMBOUT
CRUISE
DESCENT
APPROACH LANDING
TIME
MISSION SCENARIO
CA 1458 17

RELATED ISSUES

CREW DUTY TIME

REST REQUIREMENTS

DESYNCHRONOSIS

SLEEP LOSS

FATIGUE

CA1458 20

DESIRABLE OUTCOMES OF PROGRAM

MEASURE(S) WITH ALL ATTRIBUTES PREVIOUSLY LISTED

METHODS FOR ACCOMMODATING CHANGE EASILY

- IMPROVEMENTS IN MEASUREMENT METHODS
- CHANGING JOB DEMANDS

STANDARDS FOR ACCEPTABLE WORKLOAD

- HIGH WORKLOAD
- LOW WORKLOAD

CA1458.22

SOME DIFFICULT ISSUES

CERTIFICATION OF DERIVATIVES

CREW CROSS-CERTIFICATION/COMMON TYPE RATING

CONSISTENCY OF APPLICATION

EFFECTS OF TRAINING/EXPERIENCE ON WORKLOAD

STANDARDIZATION VERSUS FLEXIBILITY FOR INNOVATION AND IMPROVEMENT

CA1450 21

Intentionally left blank.

Intentionally left blank.

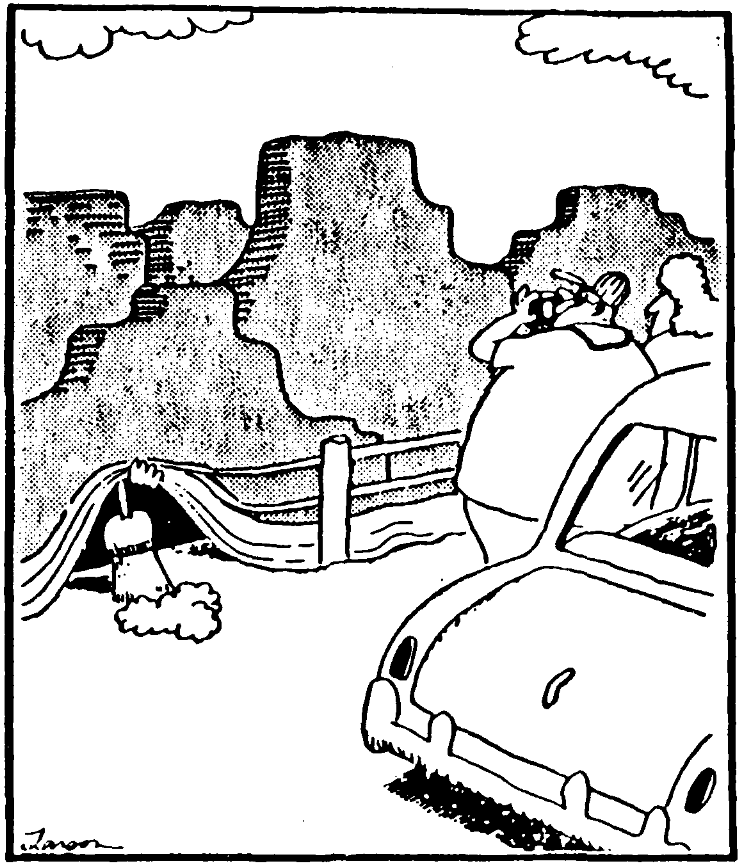
METHODOLOGICAL PROBLEMS ASSOCIATED WITH OBTAINING VALID AND RELIABLE MEASURES OF WORKLOAD DURING AIRCRAFT CERTIFICATION

BY

MICHAEL BIFERNO
DOUGLAS AIRCRAFT COMPANY
MCDONNELL DOUGLAS CORPORATION
LONG BEACH, CALIFORNIA

CA1449 04

VALIDITY — ARE YOU MEASURING WHAT YOU THINK YOU ARE MEASURING?



LARSON, 1987

VALIDITY AND RELIABILITY

VALIDITY ". . . concerns what a test measures and how well it does so." (Anastasi)

"No test (or measure) can be said to have "high" or "low" validity in the abstract. Its validity must be determined with reference to the particular use for which the test (or measure) is being considered." (Anastasi)

RELIABILITY - "Common synonyms for reliability include dependability, consistency, and stability." (Guilford)

"Despite optimum testing conditions, however, no test is a perfectly reliable instrument. Hence, every test should be accompanied by a statement of its reliability." (Anastasi)

CA1449 05

VALIDITY

PRINCIPAL CATEGORIES (APA, 1966)

- CONSTRUCT (SENSITIVITY TO VARIOUS TYPES)
- CONTENT (TYPES OF MENTAL OR PHYSICAL WORKLOAD)
- CRITERION-RELATED (PREDICTIVE)

MAJOR VALIDITY CONCERNS

- INTERNAL (CONFOUNDING)
- EXTERNAL (GENERALIZABILITY)
- DISCRIMINANT/CONVERGENT (DIFFERENTIALLY/COMMONLY SENSITIVE)
- FACE (LOOKS VALID)

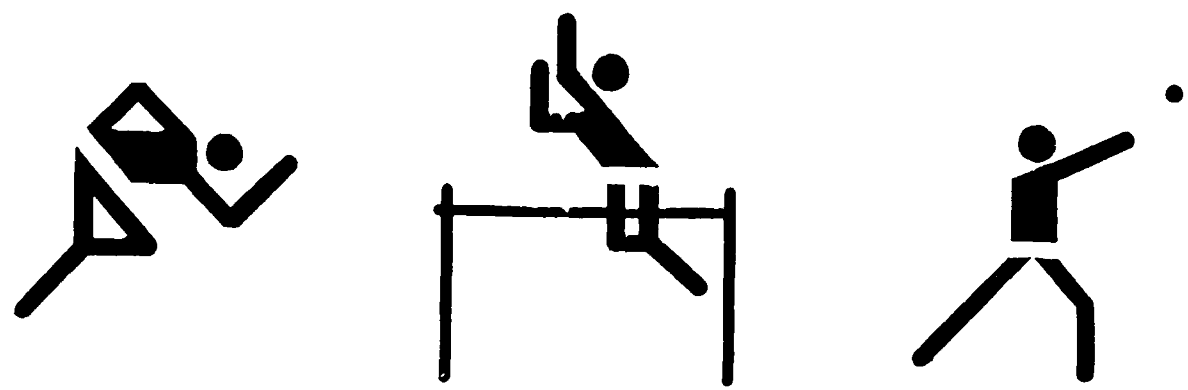
CA1449 21

CONSTRUCT VALIDITY

(Example: Best Human Athlete)

" . . . is the extent to which the test may be said to measure a theoretical construct . . . " (Anastasi)

" . . . construct validity is a comprehensive concept, which includes the other types. All specific techniques for establishing content and criterion-related validity . . . could be listed again under construct validity." (Anastasi)

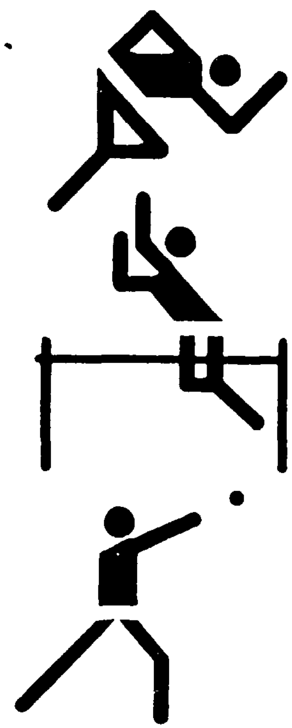


CA1449 03

CONTENT VALIDITY

(Example: Best Human Athlete)

". . . determine whether it (the test) covers a representative sample of the behavior domain to be measured." (Anastasi)



DECATHLON EVENT	SCORE
100-m dash	x
400-m dash	x
1,500-m run	x
110-m hurdles	x
Long jump	x
High jump	x
Pole vault	x
Javelin throw	x
Shot put	x
Discus throw	x
TOTAL	x,xxx

CA1449 01

CRITERION-RELATED VALIDITY
(PREDICTIVE)

(EXAMPLE: BEST HUMAN ATHLETE)

TOTAL DECATHLON SCORE → 100-METER DASH TIME

OR

400-METER DASH .IME → 100-METER DASH TIME

OR

PAST 100-METER DASH TIMES → 100-METER DASH TIME

CA1449.06

CRITERION-RELATED VALIDITY
(PREDICTIVE)

(EXAMPLE: PHYSICAL AND MENTAL WORKLOAD)

HUMAN MODEL ESTIMATES ⟶ TIME-LINE ANALYSIS RESULTS

OR

SUBJECTIVE MEASURES ⟶ TIME-LINE ANALYSIS RESULTS

OR

EMPIRICAL MEASURES ⟶ TIME-LINE ANALYSIS RESULTS

CA1449 07

CONTENT VALIDITY

(TYPES OF MENTAL OR PHYSICAL WORKLOAD)

TYPES OF MENTAL/PHYSICAL WORKLOAD PRESENT?

- FLIGHT PATH CONTROL
- COLLISION AVOIDANCE
- NAVIGATION
- COMMUNICATIONS
- OPERATIONS AND MONITORING
- COMMAND DECISIONS

IMPACT ON SCENARIO DEVELOPMENT

PROBLEM OF SEPARATING WORKLOAD TYPES

(FAR 25.1523, APPENDIX D)

CA1449 09

CONSTRUCT VALIDITY

(SENSITIVITY TO VARIOUS TYPES)

DIFFERENCES, NOT ABSOLUTE WORKLOAD LEVELS

SENSITIVITY (PROBLEMS OF UNDERLOAD)

CONVERGENT/DIVERGENT SENSITIVITY

UNITARY MEASURES VERSUS A BATTERY

CA 1449 08

CRITERION-RELATED VALIDITY

(PREDICTIVE)

ROLE OF TIME-LINE ANALYSIS IN CERTIFICATION

PREDICTION OR DIAGNOSIS?

- PREDICTION - WORKLOAD MEASURE RELATED TO FUTURE OUTCOMES
- DIAGNOSIS - WORKLOAD MEASURES RELATED TO EXISTING CONDITIONS

CA1449 10

INTERNAL VALIDITY
(LOGICAL CORRECTNESS OF CONCLUSIONS)

SOURCE OF CONFOUNDING (FACTORS JEOPARDIZING INTERNAL VALIDITY)

- HISTORY/ORDER EFFECTS/FATIGUE
- TESTING/LEARNING
- CHANGES IN INSTRUMENTATION OR OBSERVERS
- GROUP SELECTION BASED ON EXTREME SCORES
- DIFFERENTIAL LOSS OF SUBJECTS

PROBLEM OF EXPERIMENTER CONTROL OVER PILOT-SELECTED BEHAVIOR

UNCONTROLLED VARIABLES

- DIFFERENCES IN SKILL
- PERSONAL EVENTS
- MOTIVATION

NUMBER OF SUBJECTS

CA1449 11

EXTERNAL VALIDITY

(GENERALIZABILITY OF CONCLUSIONS)

THIS EFFECT CAN BE GENERALIZED TO WHICH POPULATIONS, SETTINGS, OR MEASURES?

SOME FACTORS JEOPARDIZING EXTERNAL VALIDITY

- FACE VALIDITY
- PILOT ACCEPTANCE OF METHODS
- KEY MOTIVATIONAL VARIABLES ARE ABSENT
- REACTIVE EFFECTS OF "BEING TESTED" (INTRUSION)
- MULTIPLE TREATMENT INTERFERENCE
- INTERACTION OF SELECTION BIAS AND EXPERIMENTAL VARIABLES

DISCRIMINANT/CONVERGENT VALIDITY

TRADE OFF EXTERNAL/INTERNAL VALIDITY

RELIABILITY

TEST-RETEST

ALTERNATE FORMS

INTER-RATER

CA1449 13

RELIABILITY — CONSISTENCY OF ESTIMATES

TRYING TO DESCRIBE THE SIZE OF THE BIG BANG

HARRIS, 1986

TEST-RETEST RELIABILITY

(EXAMPLE: 100-METER DASH IN SECONDS)

	HIGH RELIABILITY		LOW RELIABILITY	
ATHLETE	TEST	RETEST	TEST	RETEST
1	10	9.5	10	11
2	10.5	10	10.5	10.5
3	11	10.5	11	9.5
4	11.5	11	11.5	12
5	12	11.5	12	10
6	12.5	12	12.5	11.5
	$r = 1.0$		$r = 0.2$	

CA1449 14

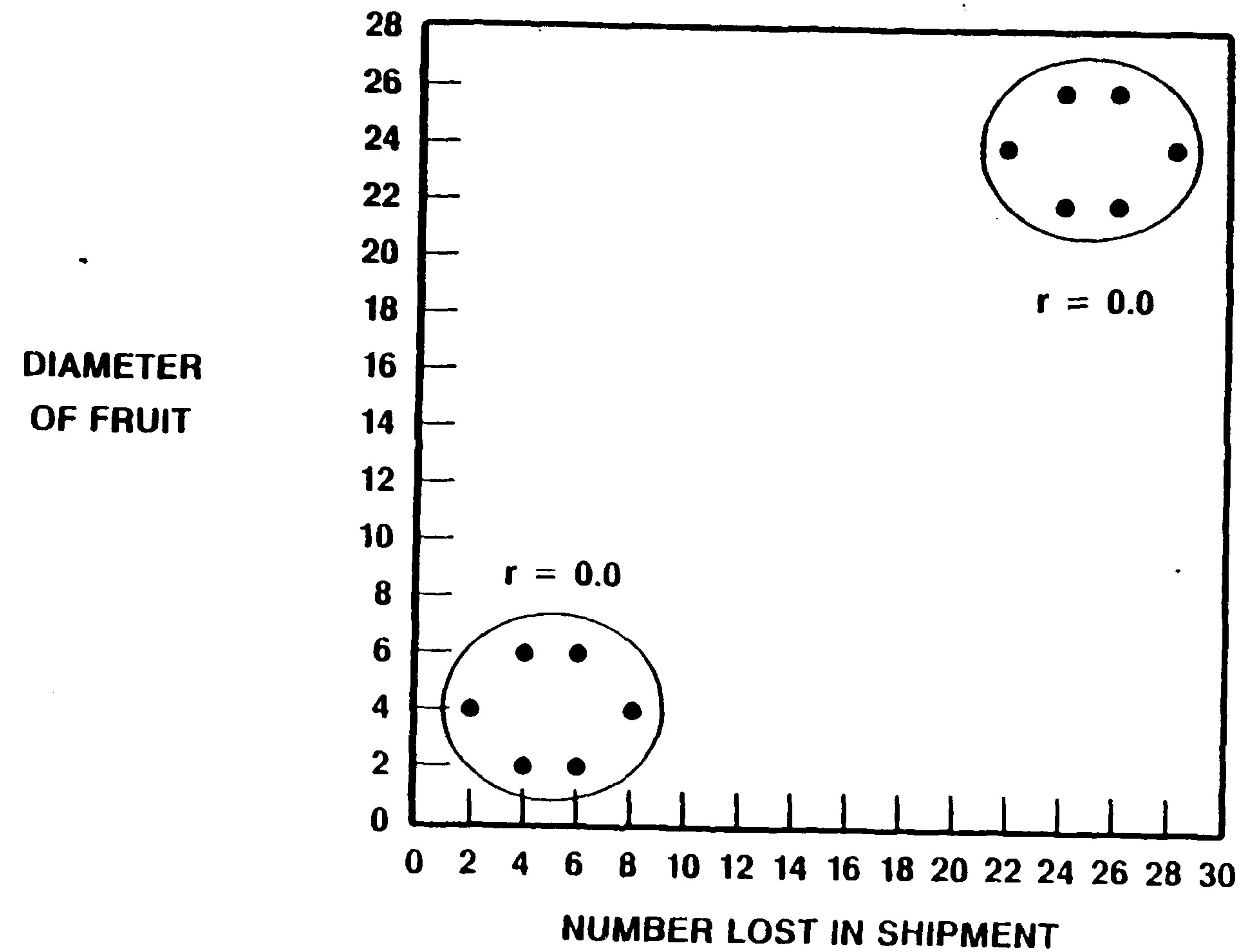
TEST-RETEST RELIABILITY

(EXAMPLE OF A BATTERY: DECATHLON SCORES)

	Test	Retest	Test	Retest
100-m dash	11 s	11 s	11 s	16 s
400-m dash	46 s	47 s	46 s	23 s
1,500-m run	240 s	245 s	240 s	360 s
110-m hurdle	14 s	13 s	14 s	21 s
Long jump	8 m	8 m	8 m	4 m
High jump	2 m	2 m	2 m	1 m
Pole vault	5 m	5 m	5 m	8 m
Javelin throw	90 m	92 m	90 m	135 m
Shot put	18 m	17 m	18 m	9 m
Discus throw	60 m	61 m	60 m	30 m
	$r = 0.99$		$r = 0.98$	

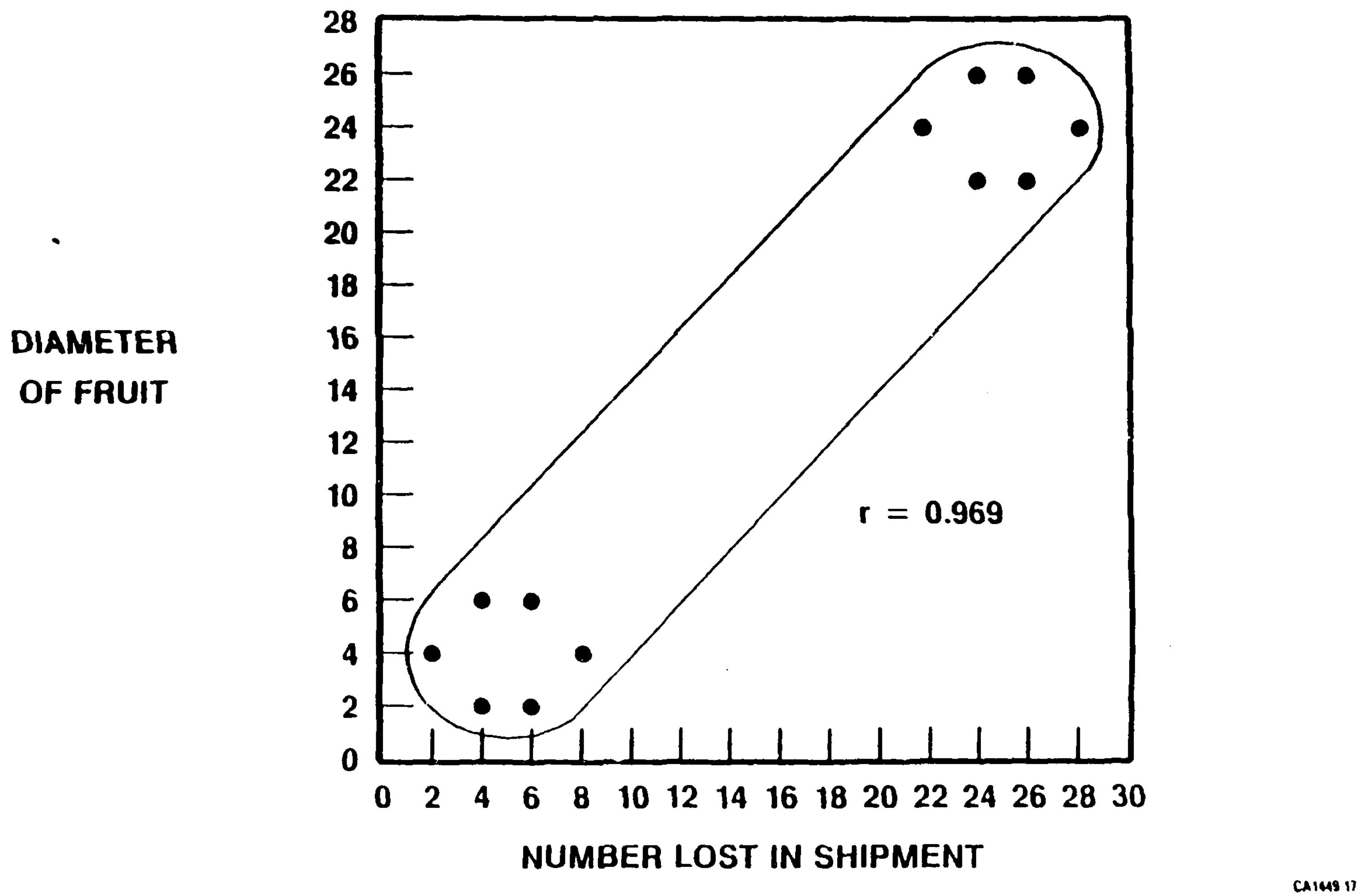
CA1449 15

APPLES AND CANTALOUPES

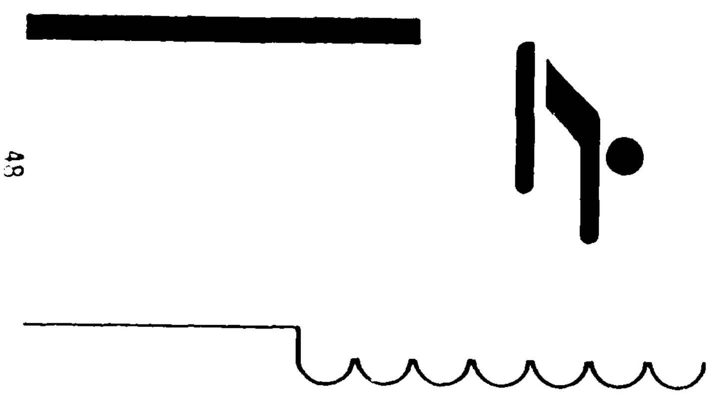


CA1449 96

APPLES AND CANTALOUPES



INTER-RATER RELIABILITY
(Example: Platform Dive)



Low Reliability

Judge	1	2	3	4	5
	8.2	9.7	8.4	8.1	6.0

High Reliability

Judge	1	2	3	4	5
	8.2	8.3	8.4	8.1	8.4

CA1449 02

TEST-RETEST/ALTERNATE FORMS RELIABILITY

CONSISTENCY OF SCORES FOR A GROUP OF PILOTS

TWO METHODS HAVE SIMILARITIES

- TEST-RETEST . . . "Answers the question concerning how stable or dependable are the measurements over time." (Guilford)
- ALTERNATE FORMS . . . "Indicates both the equivalence of content and the stability of performance." (Guilford)

CA1449 18

INTER-RATER RELIABILITY

"If raters agree (or a measure is consistent), then the number of raters (pilots) required to generate a confident decision can be reduced." (Anastasi)

CONSISTENCY OF SCORES AMONG A GROUP OF PILOTS

INDIVIDUAL DIFFERENCES

CA1440 19

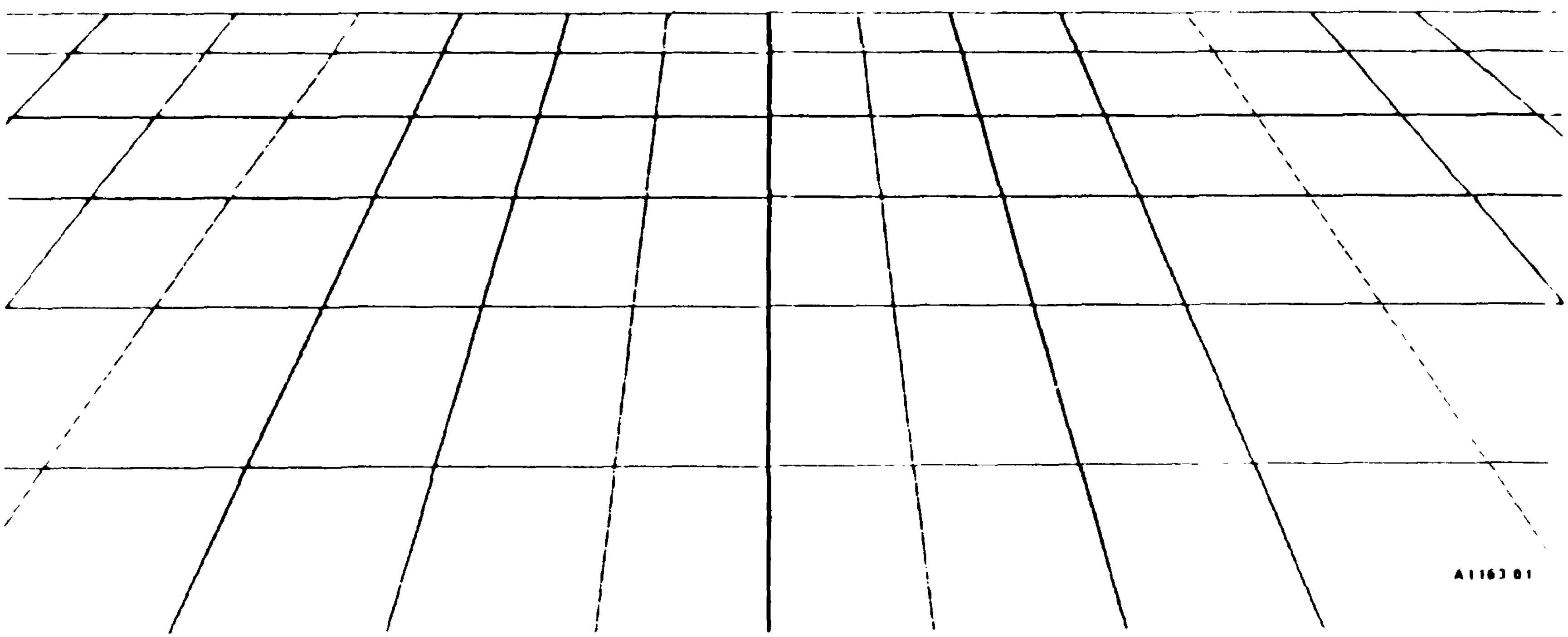
ALTERNATIVE METHODS OF ESTABLISHING RELIABILITY

REPLICATION OF RESULTS UNDER SIMILAR TEST CONDITIONS

OTHERS

CA1449 20

Practicality and Applicability Considerations for Workload Certification Measurement Techniques



A116301

Practicality

- Costs incurred
- Time contraints
- Equipment constraints

A1163 02

Costs Incurred

- Equipment costs
- Installation and preparation costs
- Time and schedule impact
- Flight and simulation costs
- Documentation costs

Time Constraints

- Certification program schedule
- Production schedules
- Delivery schedules

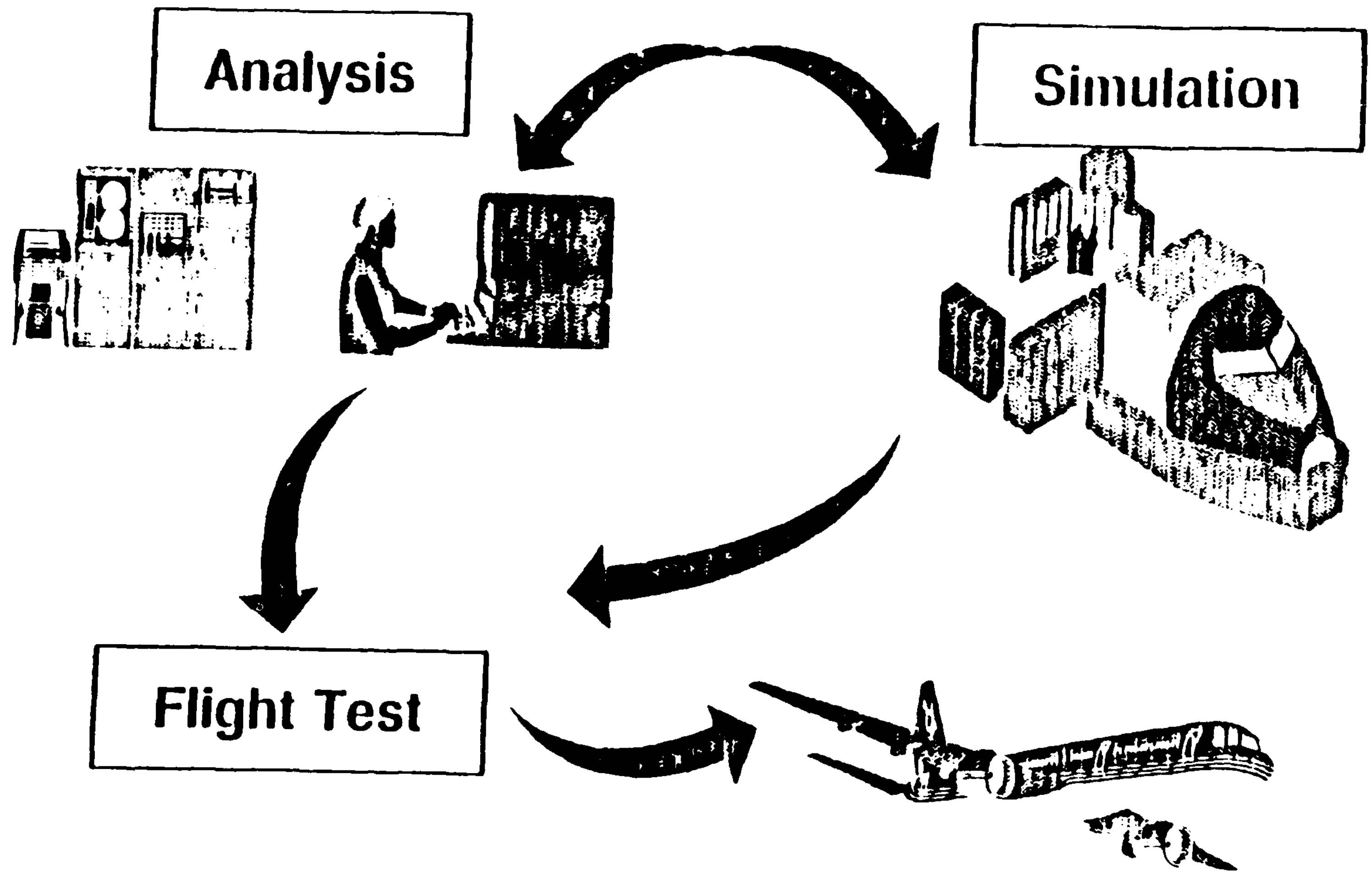
A116387

Equipment Constraints

- Limited hardware space
- Limited panel space
- Large distance between pilot and data collection hardware
- Potential signal interference
- Inability to change flight deck configuration

A1163 88

Crew Workload Assessment



A1123C 03

Applicability

- Environmental considerations
- Pilot acceptance
- Certification considerations

A1163.03

Environmental Considerations

- Must be capable of gathering data under the constraints of the flight environment

    - Part-task simulation
    - Full mission simulation
    - Flight test

- Will be applied to multidimensional task demands that could include unpredictable variations

A1163.04

Pilot Acceptance

- Nonintrusive to flight task
- Compatible with flight safety
- Compatible with normal methods of operation
- Responsive to crew self-image
- Noncareer threatening

A1163.86

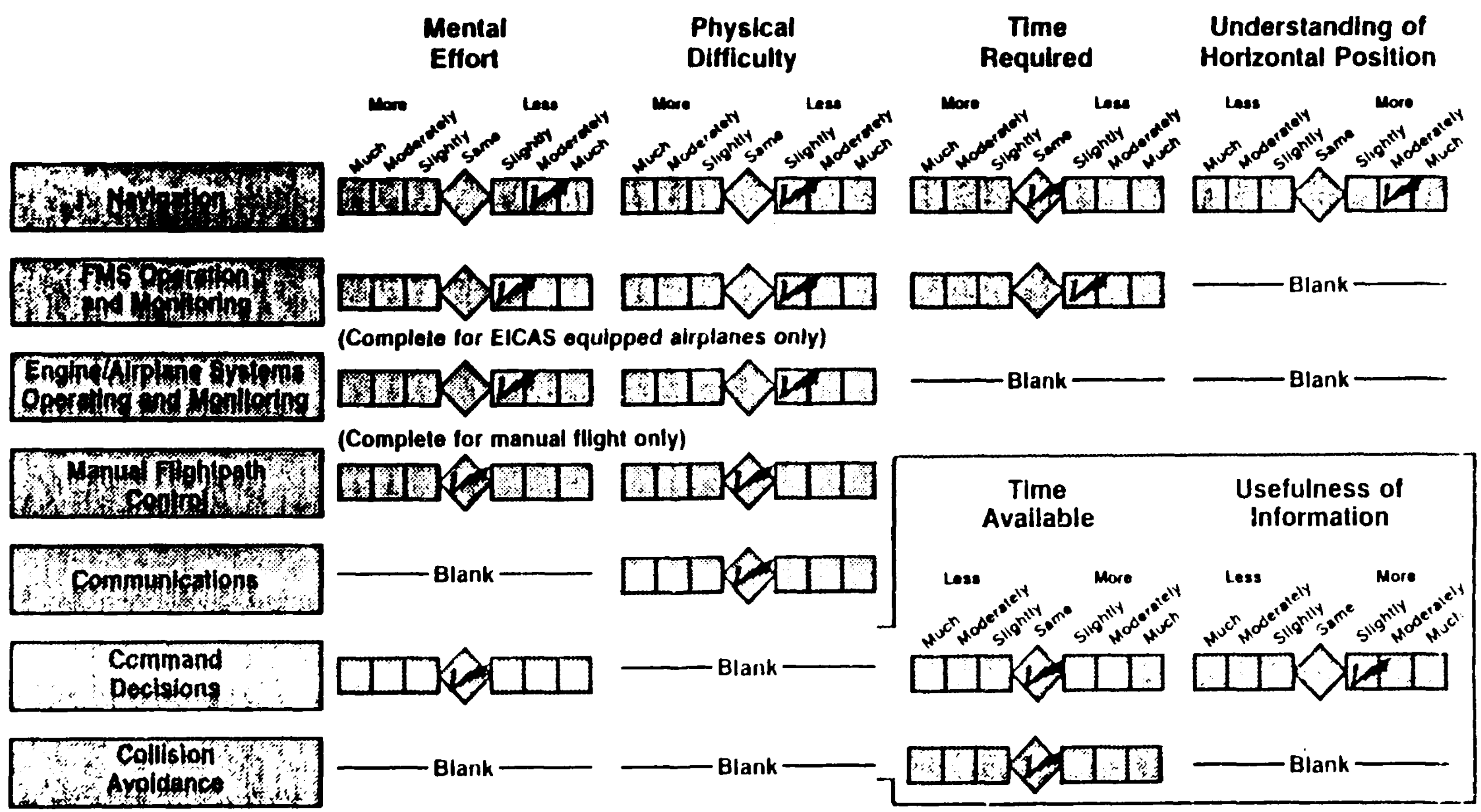
Certification Considerations

- Minimal interference with other certification flight test activities
- Technique should be appropriate for the specific phase and objectives of the certification program
- Initial efforts concentrate on those aspects that have changed from the reference aircraft

A116305

Arrival Workload Functions: Normal Operations

Compare 757/767 flightcrew operations with your reference airplane

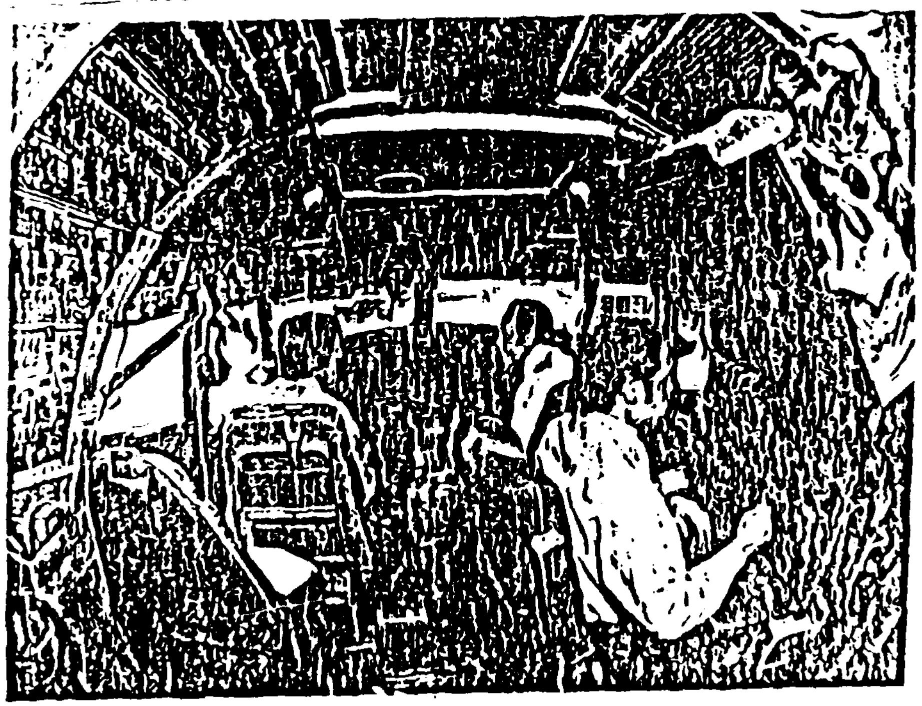


A1123.05

SIDE BY SIDE PHOTOS.

727 vs. 737 Flight Decks

727

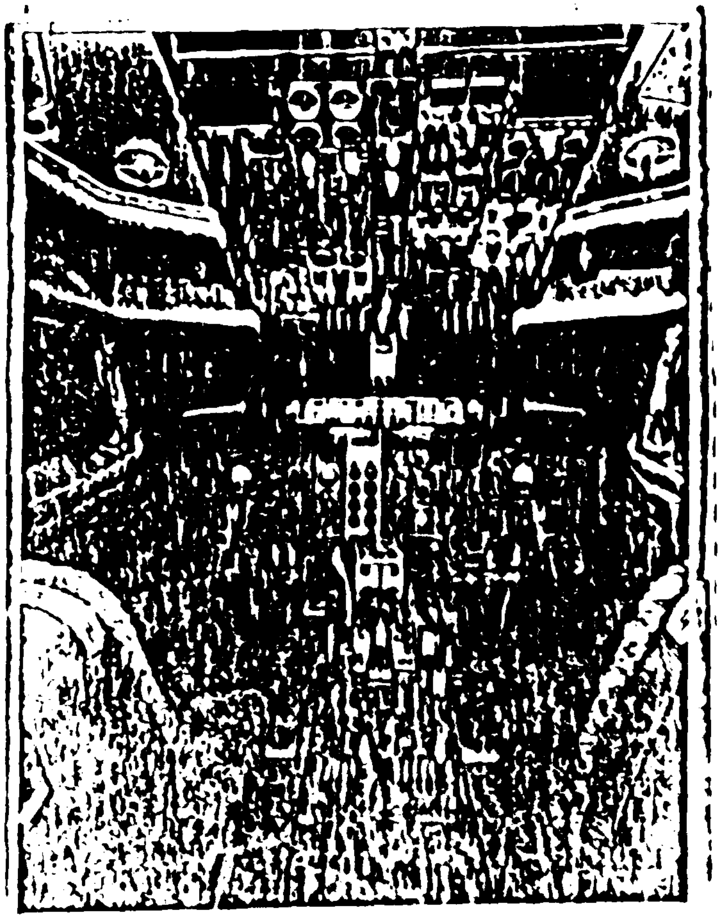


737

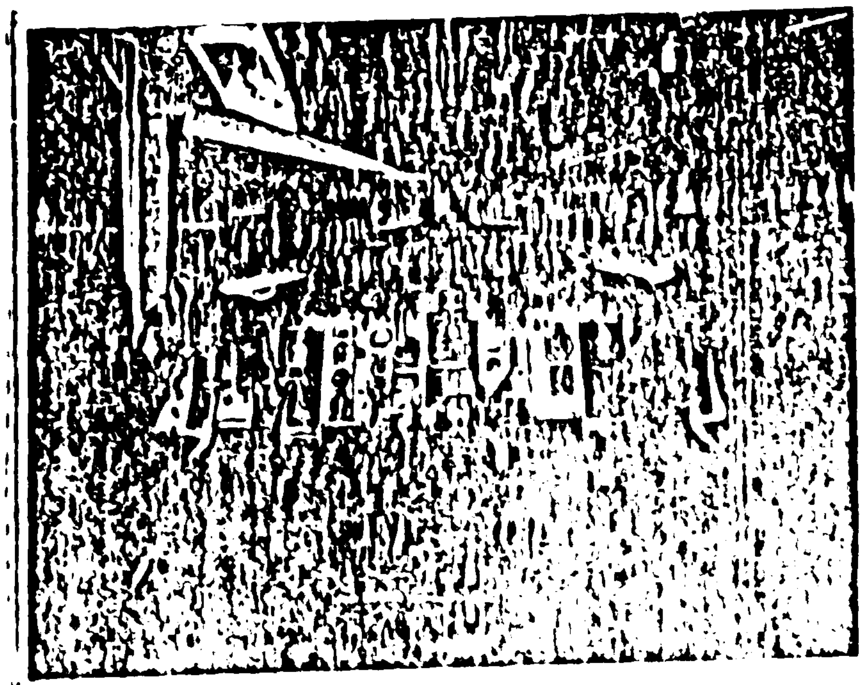
SIDE PHOTOS -

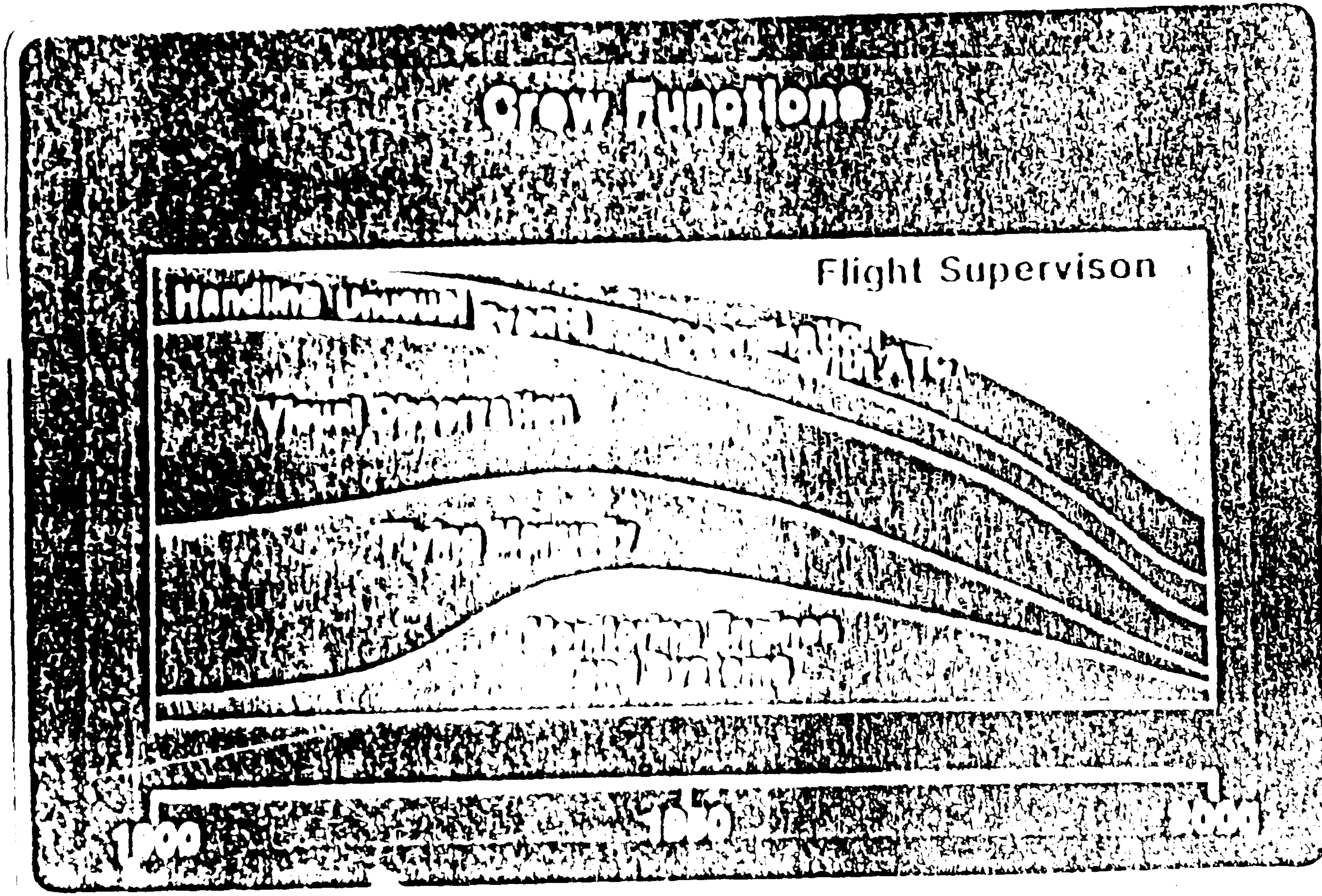
737 vs. 757/767
Flight Decks

737



757/767

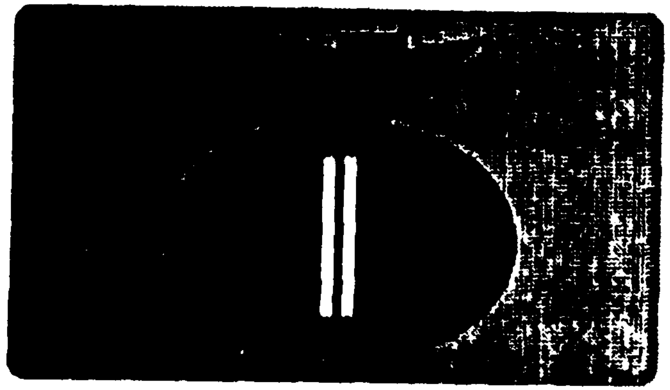



Crew Functions
Flight Supervison
Handling Unusual
1900
1950
2000

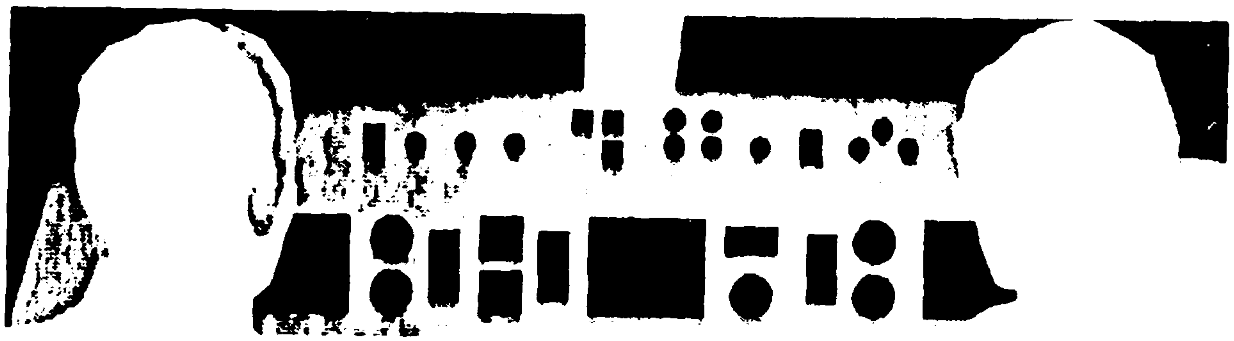
Measurement Considerations

Safety

Economics



Crew Acceptance



A1163 11

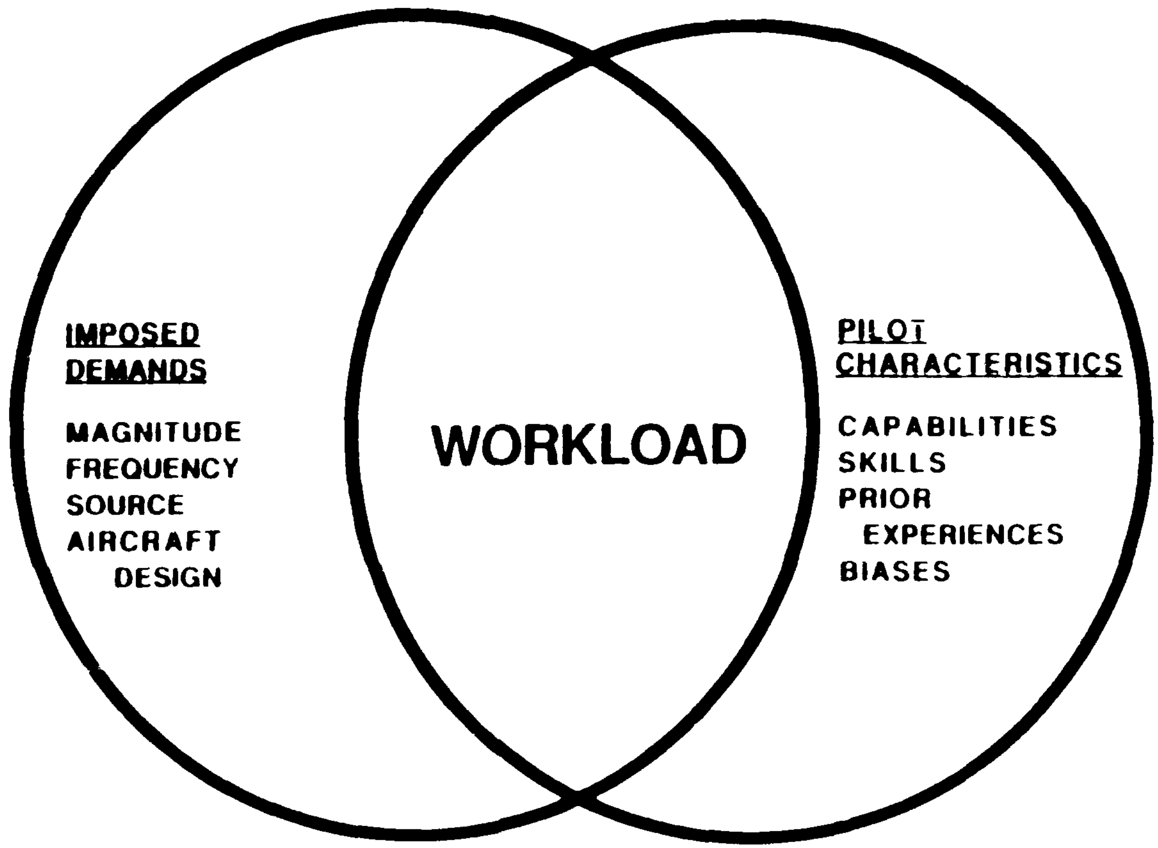
REVIEW OF SUBJECTIVE METHODS FOR MEASURING PILOT WORKLOAD

SANDRA G. HART
NASA-AMES RESEARCH CENTER
MOFFETT FIELD, CA

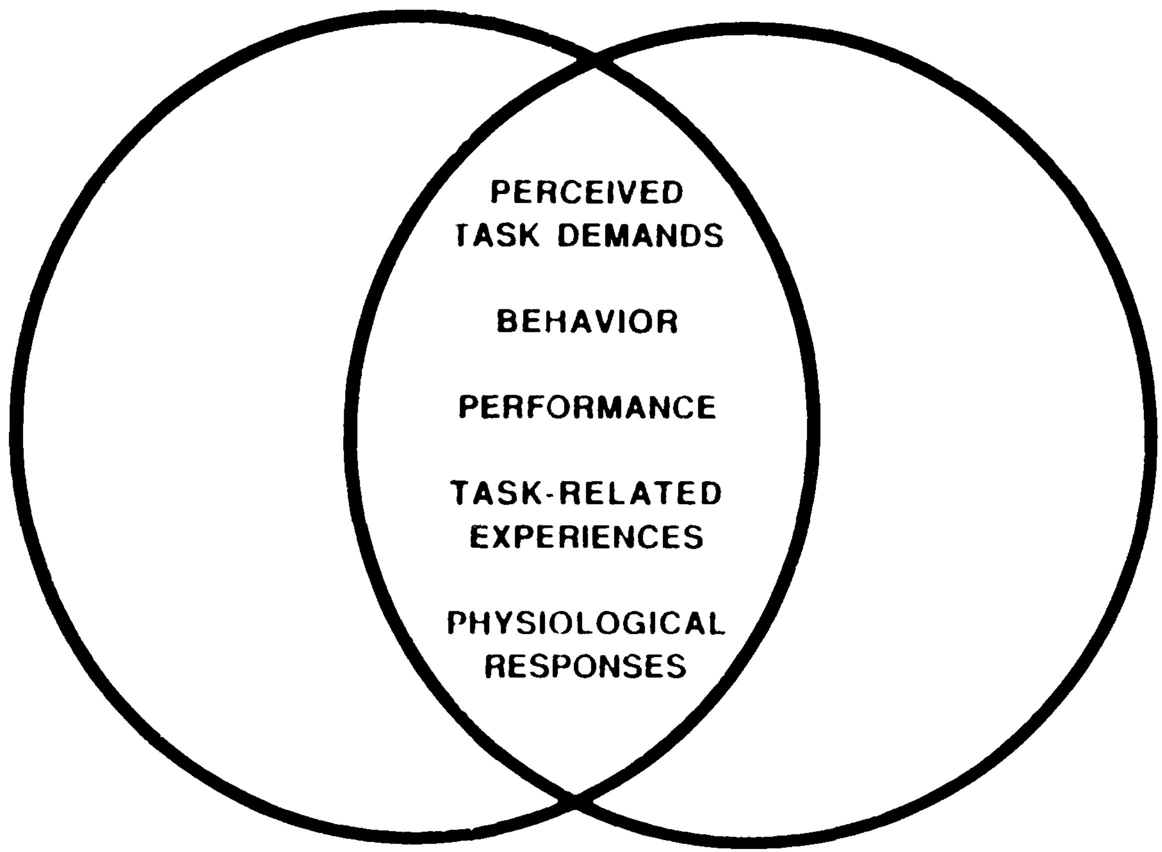
PRESENTED AT THE WORKLOAD MEASUREMENT WORKSHOP
LONG BEACH, CA
FEBRUARY 24-25, 1987

WORKLOAD: DEFINITION

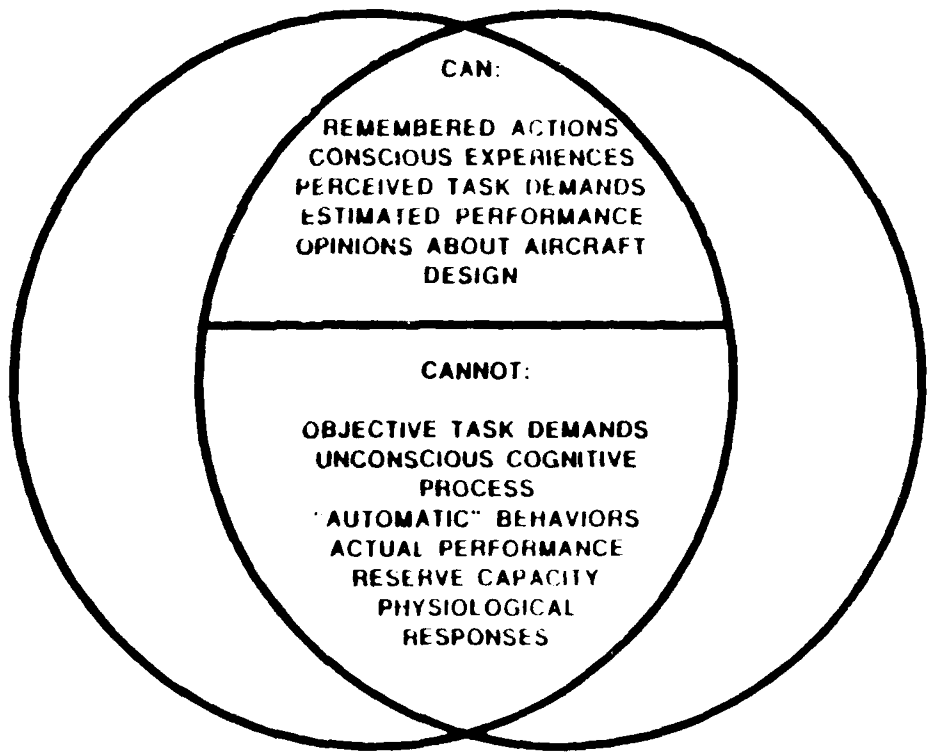
- WORKLOAD IS A HYPOTHETICAL CONSTRUCT THAT REFLECTS THE INTERACTION BETWEEN A SPECIFIC INDIVIDUAL AND THE DEMANDS IMPOSED BY A PARTICULAR TASK.
- WORKLOAD REPRESENTS THE COST INCURRED BY THE HUMAN OPERATOR IN ACHIEVING A PARTICULAR LEVEL OF PERFORMANCE



COMPONENTS OF WORKLOAD



COMPONENTS OF WORKLOAD THAT CAN (AND CANNOT) BE REPRESENTED BY SUBJECTIVE RATINGS



SUBJECTIVE RATINGS: ASSUMPTIONS

- SINCE WORKLOAD ARISES FROM THE INTERACTION BETWEEN A TASK AND THE PERFORMER, IT CANNOT BE INFERRED FROM INFORMATION ABOUT EITHER IN ISOLATION

  ...HOWEVER, TASK DEMANDS (AND THEIR ASSOCIATION WITH PERFORMANCE) OFTEN ACCOUNT FOR A SIGNIFICANT PERCENTAGE OF RATING VARIABILITY

- THERE IS SUFFICIENT COMMONALITY IN THE NATURE OF THIS INTERACTION ACROSS DIFFERENT TASKS THAT THEIR WORKLOADS CAN BE EQUATED ALONG A COMMON CONTINUUM

  ...HOWEVER, THE SPECIFIC FACTORS THAT CONTRIBUTE TO WORKLOAD VARY FROM TASK TO TASK

- DIFFERENT PEOPLE RESPOND TO AND EXPERIENCE THE SAME PHENOMENA WHEN THEY PERFORM A GIVEN TASK.

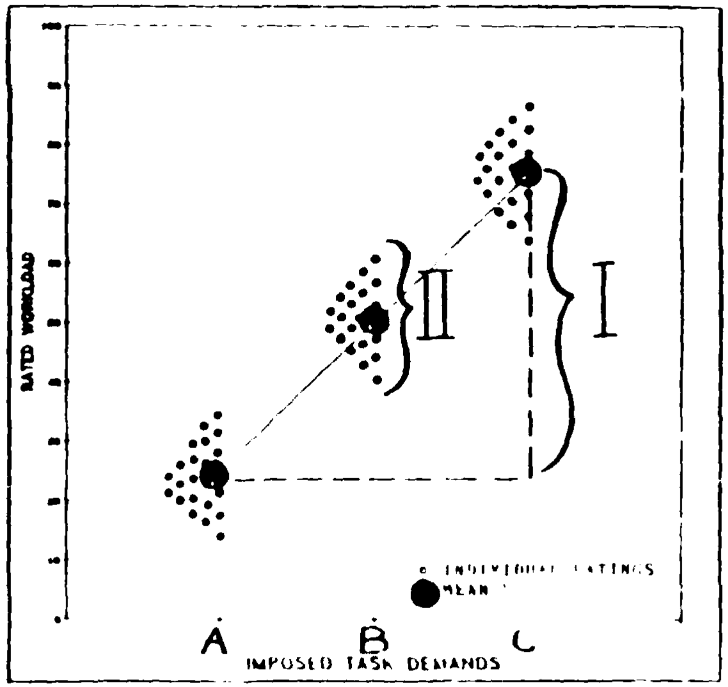
  ...HOWEVER, THERE ARE INDIVIDUAL DIFFERENCES IN SKILL, BEHAVIOR, PAST EXPERIENCE, MOTIVATION, ETC THAT INFLUENCE THE INTERACTION BETWEEN A SPECIFIC TASK AND A SPECIFIC PERFORMER

- DIFFERENT MEASURES OF WORKLOAD REFLECT THE SAME GLOBAL CONSTRUCT

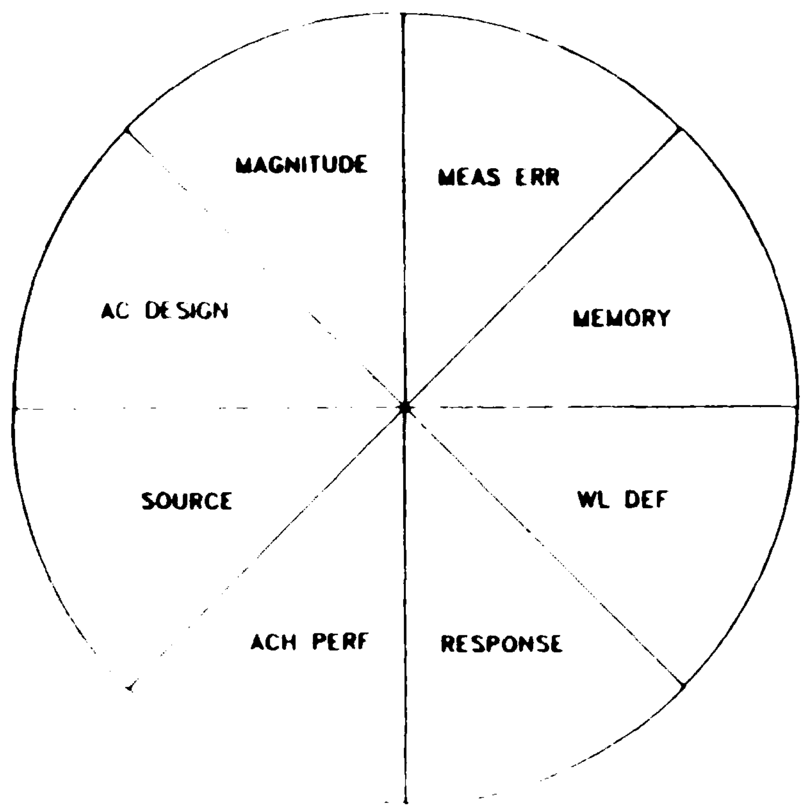
  ...HOWEVER, EACH TYPE OF MEASURE IS PARTICULARLY RESPONSIVE TO SOME COMPONENTS OF WORKLOAD AND LESS SO TO OTHERS

SUBJECTIVE RATINGS: ANALYSIS GOALS

IF A DIFFERENCE IN WORKLOAD DOES, IN FACT, EXIST AMONG TASKS (OR DISPLAYS, OR CONTROL CONFIGURATIONS, OR...) A, B, AND C, THE GOAL IS TO MAXIMIZE THE SCALE'S SENSITIVITY TO I AND TO MINIMIZE ITS SENSITIVITY TO II.

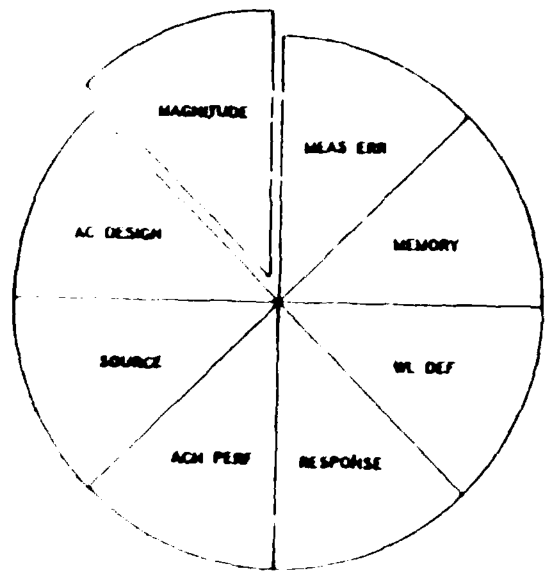


SOURCES OF VARIABILITY IN WORKLOAD RATINGS



SUBJECTIVE RATINGS: SOURCES OF VARIABILITY
INTENSITY OF IMPOSED DEMANDS

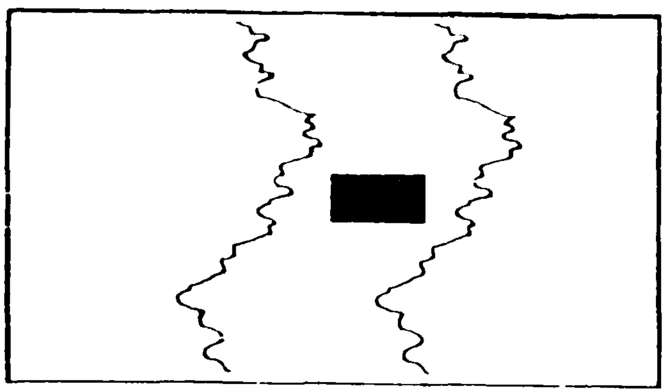
- REQUIRED ACCURACY
- REQUIRED SPEED
- COMPLEXITY
- FREQUENCY OF RESPONSE
- TIME CONSTRAINTS
- CONCURRENT SUBTASKS
- INFORMATION LOAD

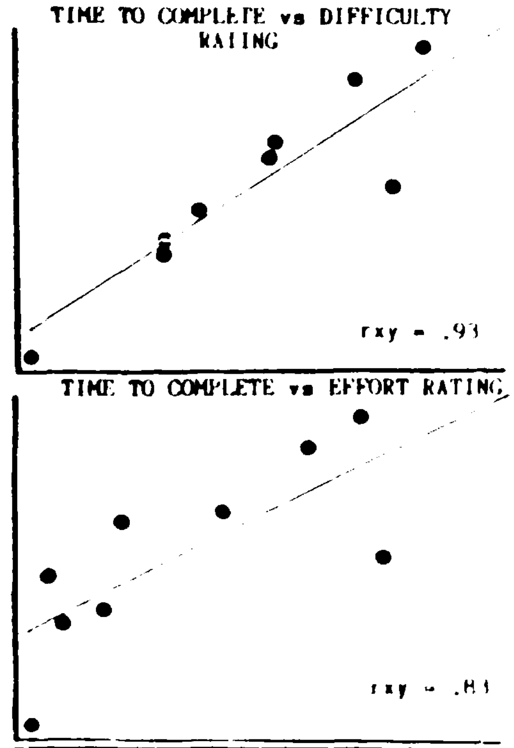


RATING SCALES: TWO UNIDIMENSIONAL SCALES (EFFORT, DIFFICULTY)

EXAMPLE: SUBJECTIVE RATINGS OF EFFORT AND TASK DIFFICULTY WERE OBTAINED ON 10-POINT SCALES DURING DIFFERENT VERSIONS OF A SIMULATED HOVERCRAFT CONTROL TASK. SKILL-BASED, RULE-BASED, AND KNOWLEDGE-BASED' ASPECTS OF WORKLOAD WERE EXPERIMENTALLY MANIPULATED

HOVERCRAFT TASK DISPLAY





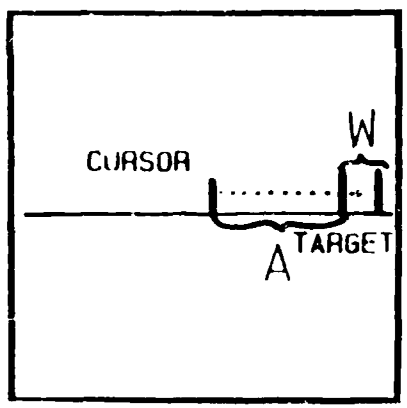
NOTE: THE GREATER THE LEVEL OF TASK DIFFICULTY, THE MORE PRONOUNCED THE DISPARITY BETWEEN ACTUAL AND PERCEIVED EFFORT.

Chan, Krushelnycky, Allen, & Discepola, 1985

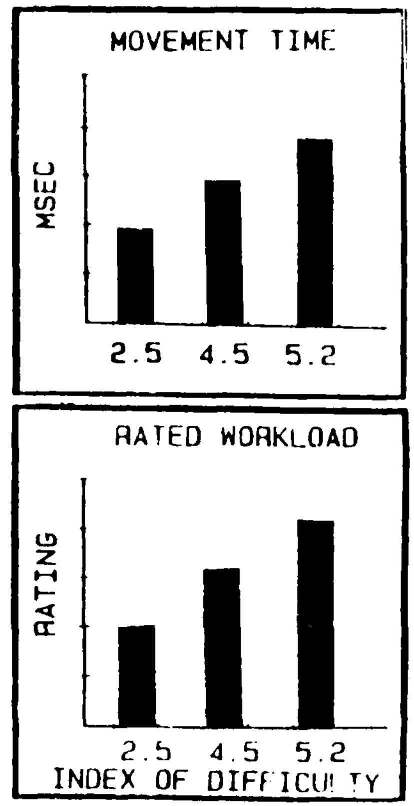
SUBJECTIVE RATING: CONSTRUCT VALIDITY

IDEALLY, RATINGS SHOULD BE EVALUATED WITH RESPECT TO THEORY BASED ESTIMATES OF IMPOSED TASK DEMAND LEVELS.

EXAMPLE. NASA-TASK LOAD INDEX RATINGS WERE COMPARED TO THE DIFFICULTY (MEASURED BY FITTS LAW) OF TARGET ACQUISITION TASKS



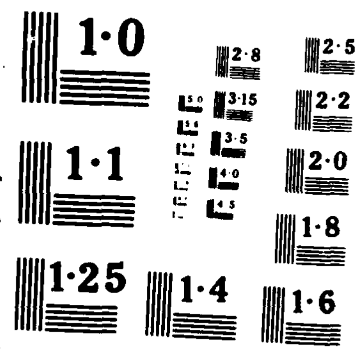
$$ID = LOG_2 \left(\frac{2A}{W}\right)$$



Yeh, Wickens & Hart, 1985

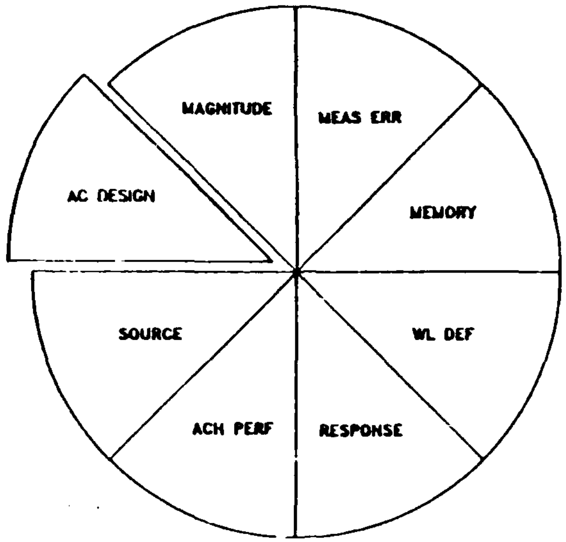
AD-A189 004 PROCEEDINGS OF THE WORKSHOP ON THE ASSESSMENT OF CREW WORKLOAD MEASUREMEN..(U) DOUGLAS AIRCRAFT CO LONG BEACH CA M A BIFERNO ET AL. JUN 87 AFWAL-TR-87-3043-VOL-1 2/4

UNCLASSIFIED F33615-86-C-3600 F/G 5/9 NL


1·0
2·8
2·5
3·15
2·2
3·5
2·0
1·1
4·0
1·8
1·25
1·4
1·6

SUBJECTIVE RATINGS: SOURCES OF VARIABILITY AIRCRAFT DESIGN

- DISPLAY DESIGN
  - FORMAT
  - INFORMATION QUALITY
  - INFORMATION QUANTITY
- CONTROL DESIGN
- HANDLING QUALITIES
- DISPLAY/CONTROL COMPATIBILITY
- AUTOMATIC SUBSYSTEMS
  - DESIGN, INTEGRATION
  - INTERFACE



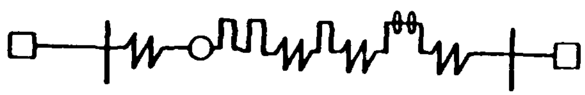
SUBJECTIVE RATINGS: POTENTIAL GOALS OF A WORKLOAD ANALYSIS:

* MAGNITUDE OF DIFFERENT TASK DEMANDS
* WORKLOAD-IMPACT OF ALTERNATIVE DESIGNS
* EFFECT OF CREW-COMPLEMENT ON WORKLOAD

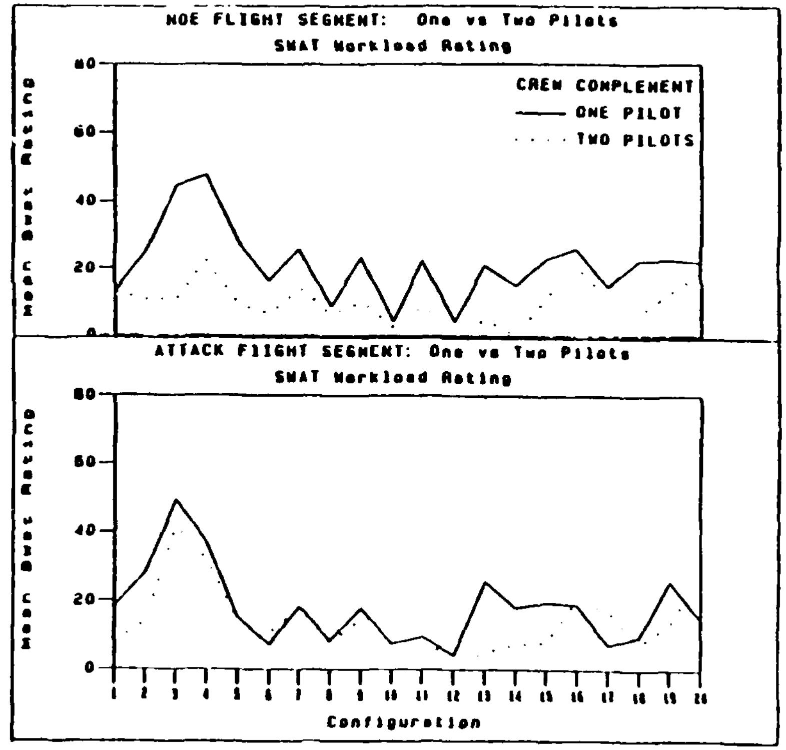
EXAMPLE: SWAT AND NASA BIPOLAR RATINGS WERE OBTAINED DURING AN ADVANCED HELICOPTER SIMULATION.

LEVEL OF AUTOMATION, CREW SIZE (ONE vs TWO), FLIGHT TASK (BOB-UP, HOVER, NOE, AIR-AIR, ETC) WERE VARIED.

MISSION SCENARIO PROFILE



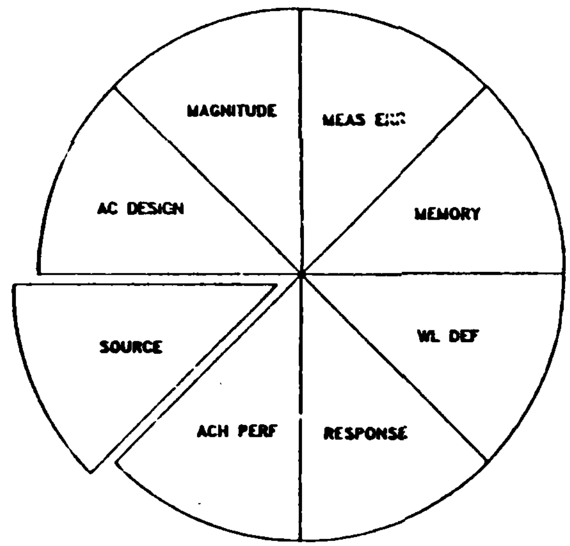
START AND END WAYPOINT
NAVIGATION
PREDETERMINED HOVER/FIRE POINT
HOVER AND BOB-UP MANEUVER
AIR TO AIR ENGAGEMENT
ELAPSED TIME/OR DISTANCE



Haworth, Bivens, & Shively, 1986

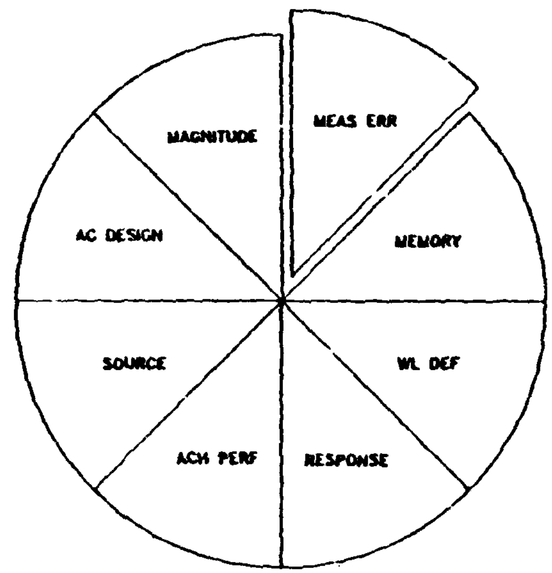
SUBJECTIVE RATINGS: SOURCES OF VARIABILITY
SPECIFIC SOURCES OF DEMANDS

- PSYCHOLOGICAL VARIABLES
  - STIMULUS MODALITY
  - RESPONSE MODALITY
  - PROCESSING STAGE
  - PROCESSING CODE
- TASK-RELATED VARIABLES
  - FLIGHT-PATH CONTROL
  - COLLISION AVOIDANCE
  - NAVIGATION
  - COMMUNICATIONS
  - SYSTEMS MONITORING
  - COMMAND DECISIONS
- ENVIRONMENTAL VARIABLES
  - SOCIAL
  - PHYSICAL

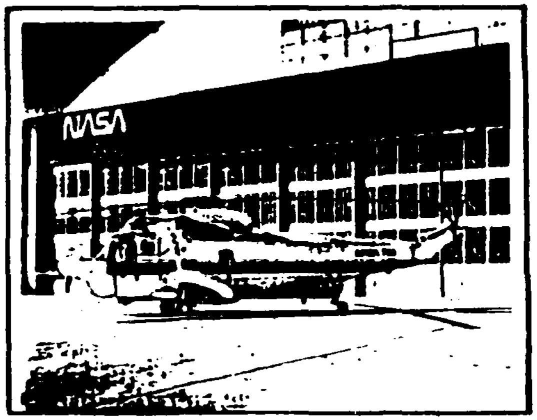


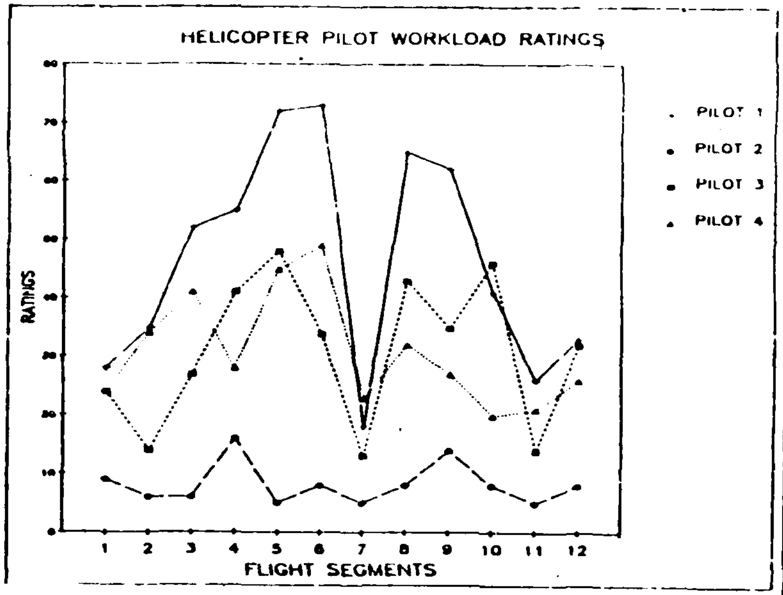
SUBJECTIVE RATINGS: SOURCES OF VARIABILITY
MEASUREMENT ERROR

- MEANINGFULNESS OF UNITS
- VERBAL DESCRIPTORS
- SEQUENTIAL BIASES
- CENTERING BIAS
- INDIVIDUAL DIFFERENCES IN USE OF SCALE
- NUMBER OF ALTERNATIVE SCALE VALUES



SUBJECTIVE RATINGS: MEASUREMENT ERROR
INDIVIDUAL DIFFERENCES IN THE USE OF A SCALE





EXAMPLE: NASA-TLX RATINGS WERE OBTAINED DURING EXPERIMENTAL FLIGHTS OF AN SH-3G HELICOPTER. ALTHOUGH THE PILOTS WERE HIGHLY EXPERIENCED TEST PILOTS, AND PERFORMED EQUALLY WELL, THE ABSOLUTE LEVELS OF THEIR RATINGS VARIED CONSIDERABLY.

SUBJECTIVE RATINGS: MEASUREMENT ERROR

CONTEXT EFFECTS

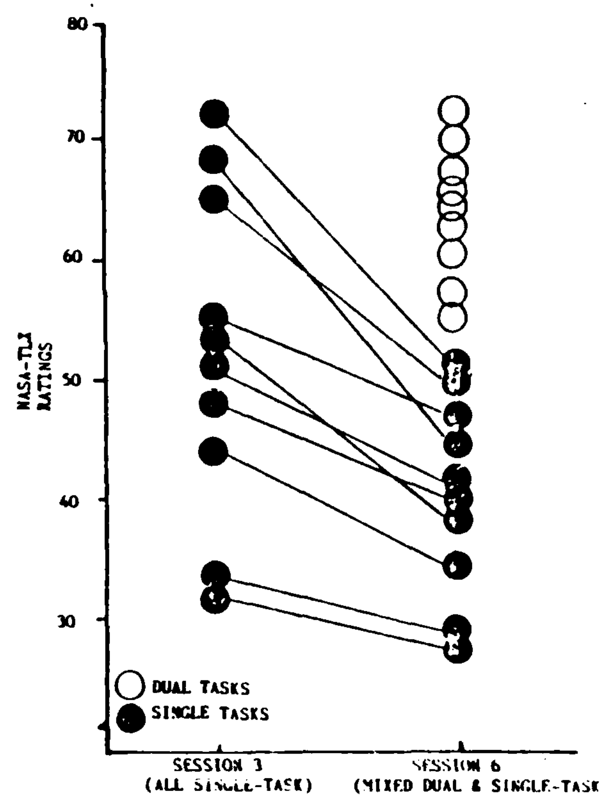
- RATERS TEND TO USE THE FULL RANGE OF A SCALE TO QUANTIFY THE WORKLOAD OF TASKS PRESENTED IN EACH CONTEXT

- THIS MAY RESTRICT THE RANGE OF RATINGS AVAILABLE FOR TASKS PRESENTED AT THE END OF THE SEQUENCE

- RATING VALUES DO NOT REPRESENT ABSOLUTE JUDGEMENTS THAT CAN BE COMPARED ACROSS CONTEXTS, BUT RELATIVE COMPARISONS WITHIN EACH CONTEXT

- EXAMPLE:
  NASA-TLX RATINGS WERE OBTAINED AFTER PERFORMANCE OF A SERIES OF SINGLE-AXIS TRACKING TASKS AND AGAIN WHEN THE SAME SINGLE-AXIS TASKS WERE PRESENTED INTERSPERCED WITH DUAL-AXIS TRACKING TASKS

- RESULTS:
  THE RELATIVE ORDERING OF SINGLE-TASK RATINGS DID NOT CHANGE SIGNIFICANTLY, BUT THE RANGE OF SINGLE TASK RATINGS WAS RESTRICTED.



Vidulich & Tsang (in press)

SUBJECTIVE RATINGS: MEASUREMENT ERROR

CONTEXT EFFECTS

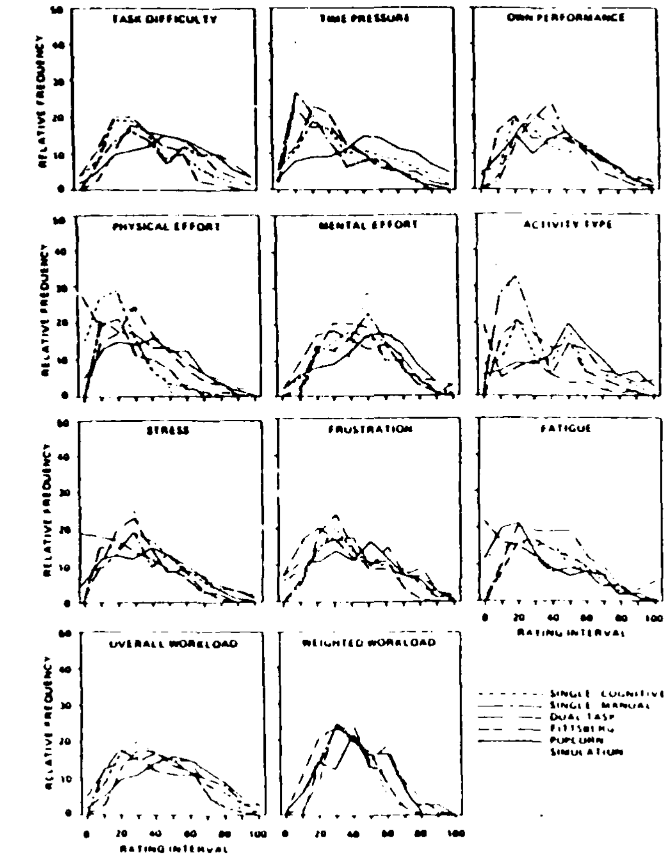
- RATERS TEND TO USE THE FULL RANGE OF A SCALE TO QUANTIFY THE WORKLOAD OF TASKS PRESENTED IN EACH CONTEXT

- THUS, ABSOLUTE MAGNITUDES OF THE RATINGS ARE NOT MEANINGFUL ACROSS DIFFERENT CONTEXTS, ALTHOUGH THE RANK ORDERING AND RELATIVE DISTANCES BETWEEN THE RATINGS DO PROVIDE USEFUL INFORMATION WITHIN EACH CONTEXT

- EXAMPLE:
  NASA-BIPOLAR RATINGS AND GLOBAL WORKLOAD RATINGS WERE OBTAINED DURING 16 EXPERIMENTS. THEY WERE GROUPED INTO SIX CATEGORIES OF ACTIVITIES.

- RESULTS:
  THE DISTRIBUTIONS OF RATINGS WERE SURPRISINGLY SIMILAR ACROSS TASK CATEGORIES THAT APPEAR TO DIFFER WIDELY IN DIFFICULTY

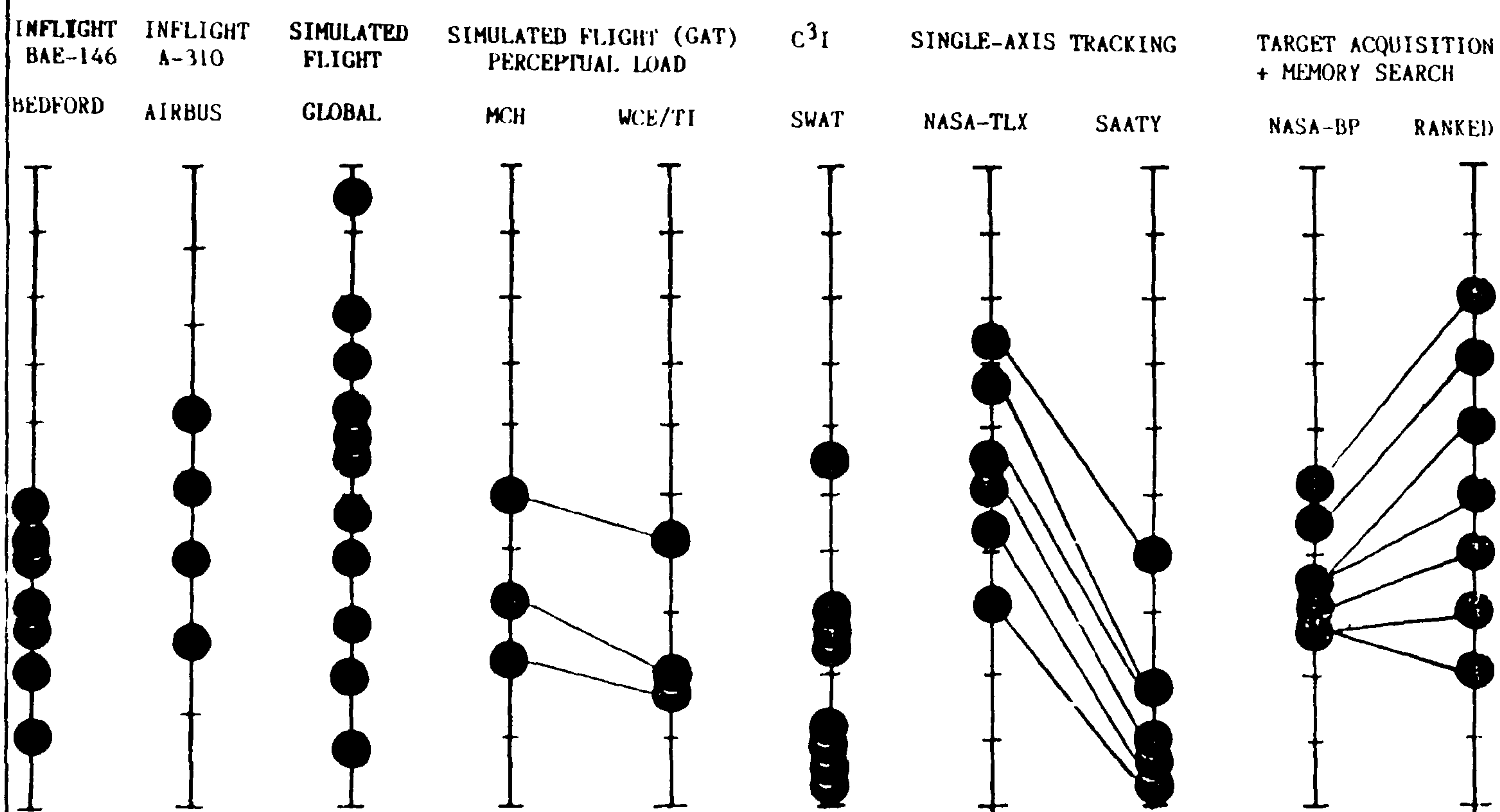
  CONSISTENTLY SIGNIFICANT DIFFERENCES IN RATINGS WERE FOUND WITHIN EACH CATEGORY IN RESPONSE TO EXPERIMENTAL MANIPULATIONS



HART & STAVELAND, in press

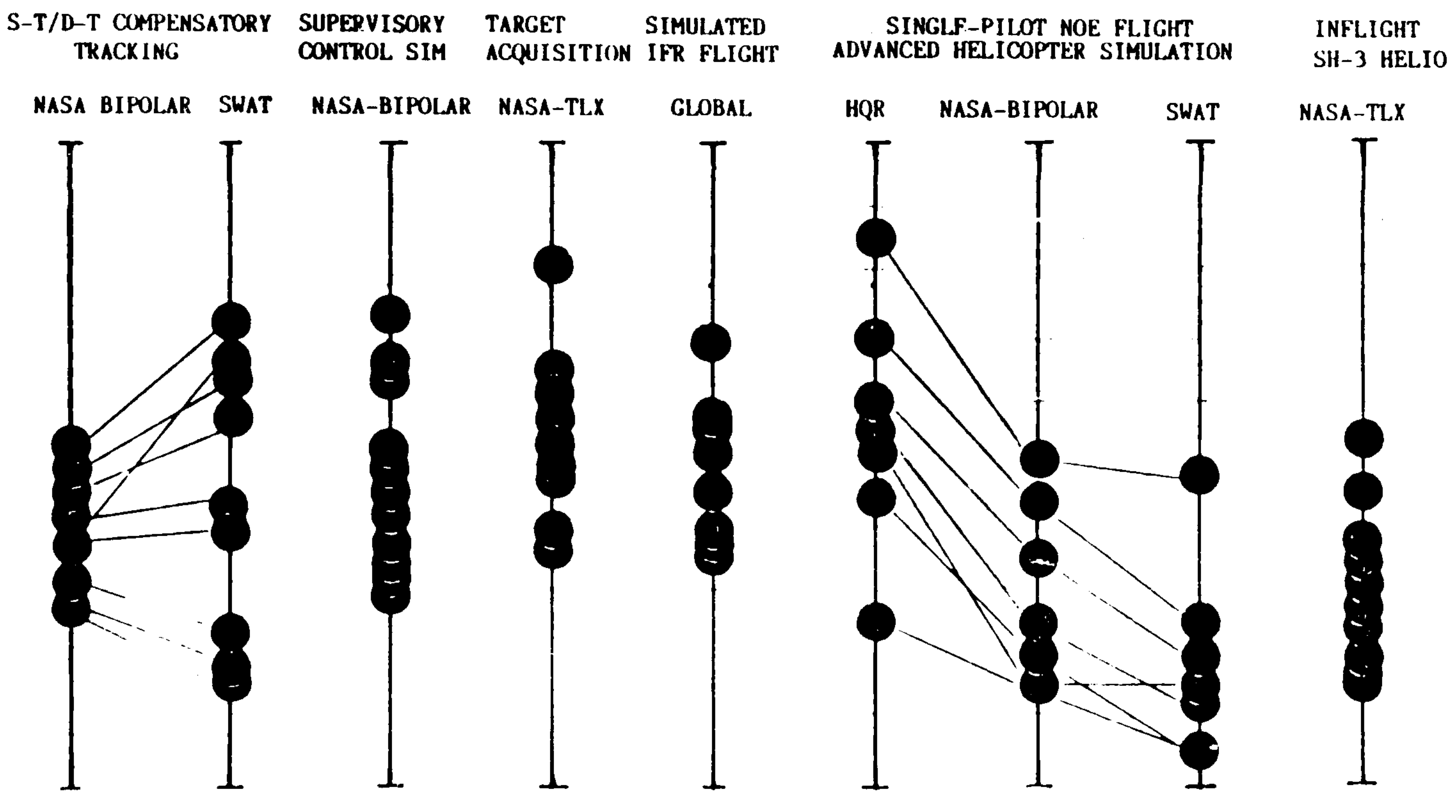
SUBJECTIVE RATINGS: MEASUREMENT ERROR CONTEXT EFFECTS

AVERAGE WORKLOAD RATINGS OBTAINED IN DIFFERENT EXPERIMENTS (CONT):



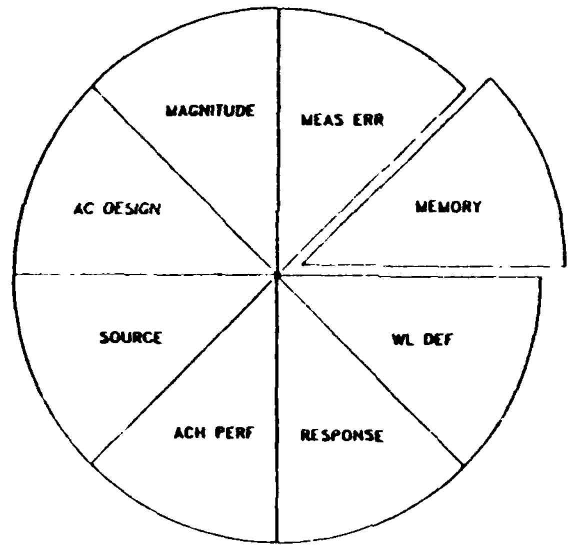
SUBJECTIVE RATINGS: MEASUREMENT ERROR CONTEXT EFFECTS

AVERAGE WORKLOAD RATINGS OBTAINED IN DIFFERENT EXPERIMENTS USING:



SUBJECTIVE RATINGS: SOURCES OF VARIABILITY MEMORY

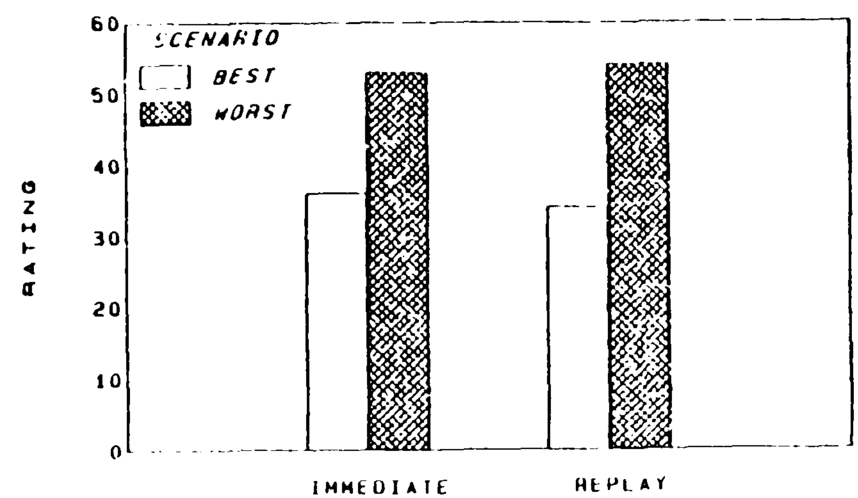
- HOMOGENEITY OF TASKS WITHIN INTERVAL RATED
- CONSCIOUS AWARENESS OF RELEVANT FACTORS
- TIME SINCE PRESENTATION OF REFERENCE TASK
- TIME SINCE INTERVAL ENDED
- MNEMONIC AIDES



SUBJECTIVE RATINGS: MEMORY

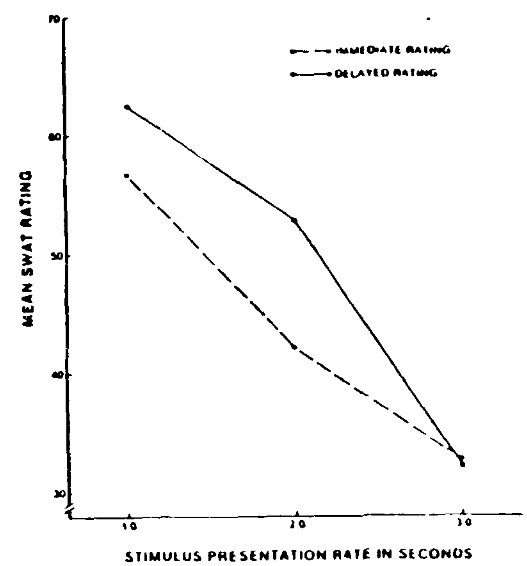
SINCE RATINGS REPRESENT THE RATER'S MEMORY OF WHAT WAS EXPERIENCED DURING AN INTERVAL, THEY SHOULD BE OBTAINED IMMEDIATELY. HOWEVER, NASA BIPOLAR RATINGS OBTAINED IMMEDIATELY AND AFTER A DELAYED REPLAY OF THE INTERVALS TO BE EVALUATED (A) AND SWAT RATINGS OBTAINED IMMEDIATELY OR AFTER A DELAY OF 15 MIN (B) WERE NOT SIGNIFICANTLY CHANGED BY THE DELAY.

a. NASA BIPOLAR RATING SCALE



Hart, et al, in press

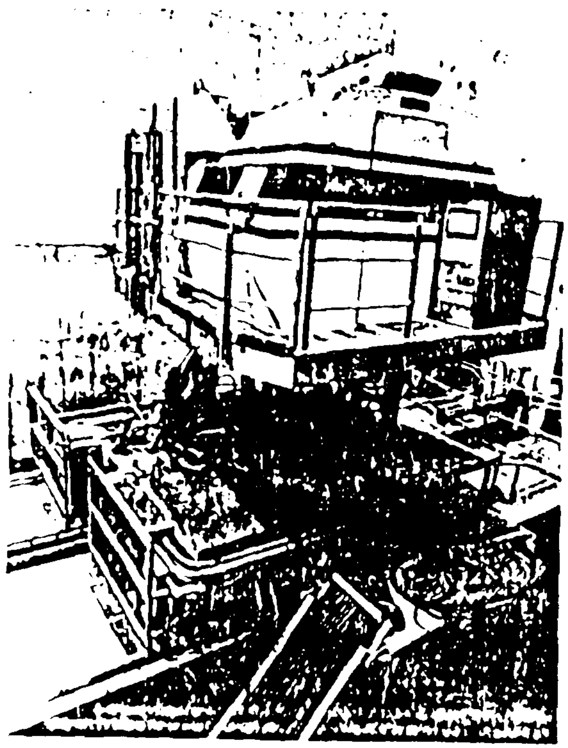
b. SWAT



Eggemeier, Crabtree & LaPointe 1983

SUBJECTIVE RATINGS: MEMORY

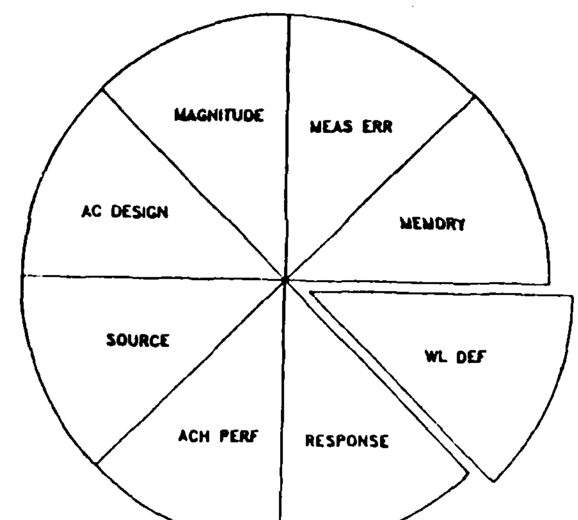
SINCE RATINGS REPRESENT THE RATER'S MEMORY OF WHAT WAS EXPERIENCED DURING AN INTERVAL, THEY SHOULD BE OBTAINED IMMEDIATELY. HOWEVER, SWAT RATINGS OBTAINED IMMEDIATELY AFTER EACH FLIGHT SEGMENT AND NASA BIPOLAR RATINGS OBTAINED AFTER A DELAYED VIDEO TAPE REPLAY OF THE SIMULATED FLIGHTS, WERE HIGHLY CORRELATED.



Haworth, Bivens, & Shively, 1986

SUBJECTIVE RATINGS: SOURCES OF VARIABILITY
INDIVIDUAL DIFFERENCES IN DEFINITION

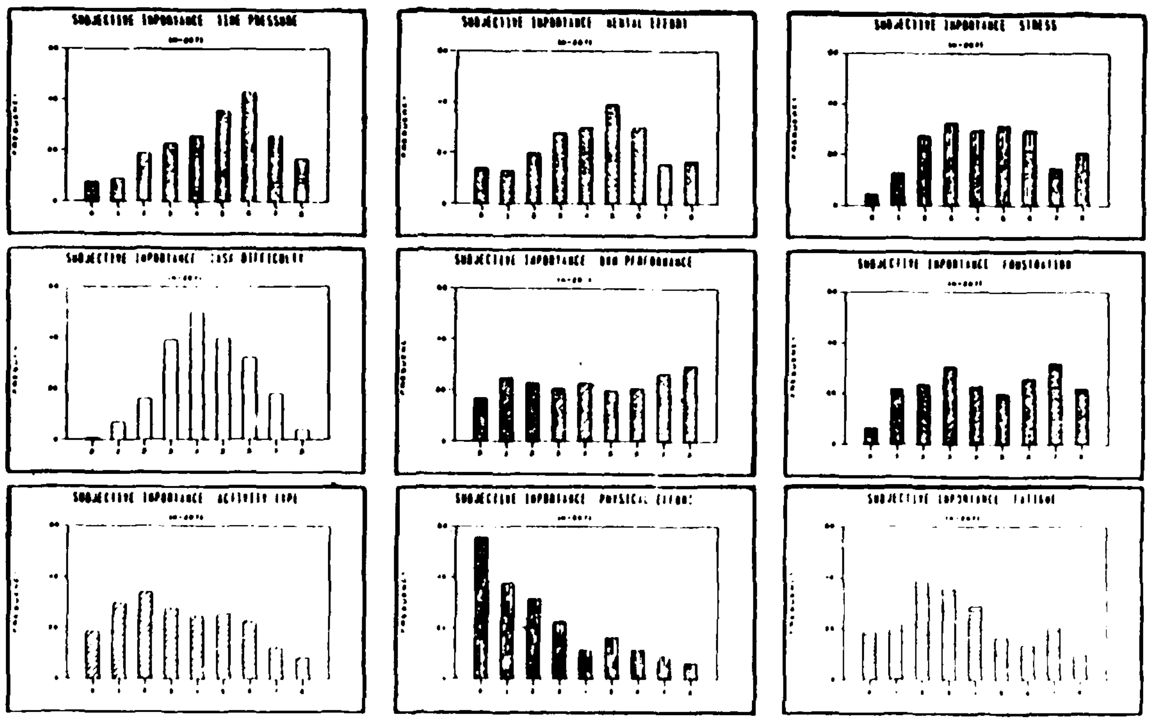
- A PRIORI BIASES
  - PHYSICAL EFFORT
  - MENTAL EFFORT
  - TIME PRESSURE
  - STRESS
  - PERFORMANCE
- SUBJECTIVE RELEVANCE OF EXPERIMENTALLY MANIPULATED SOURCES OF LOAD
- SENSITIVITY TO EACH COMPONENT OF LOAD



SUBJECTIVE RATINGS: INDIVIDUAL DIFFERENCES IN DEFINITION

THE RELATIVE IMPORTANCE OF DIFFERENT FACTORS TO EACH INDIVIDUAL'S DEFINITION OF WORKLOAD VARY CONSIDERABLY.

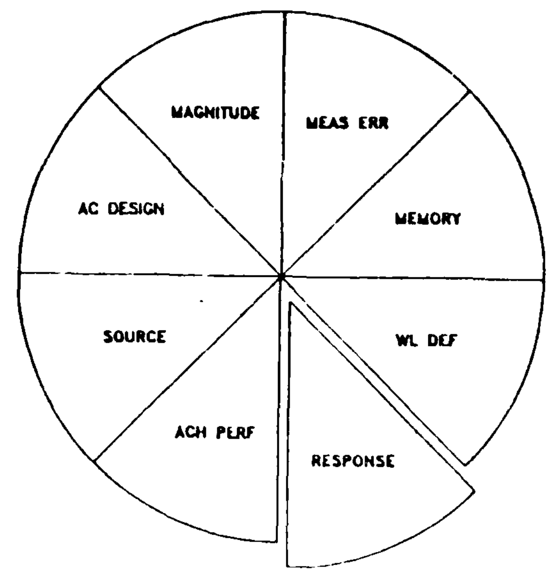
EXAMPLE: THE RELATIVE FREQUENCY DISTRIBUTIONS OF IMPORTANCE 'WEIGHTS' ATTACHED TO 9 FACTORS FOR 207 SUBJECTS ARE REPRESENTED BELOW.



Hart & Staveland, in press

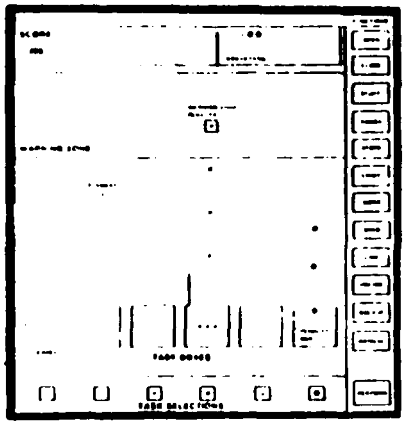
SUBJECTIVE RATINGS: SOURCES OF VARIABILITY RESPONSE TO TASK

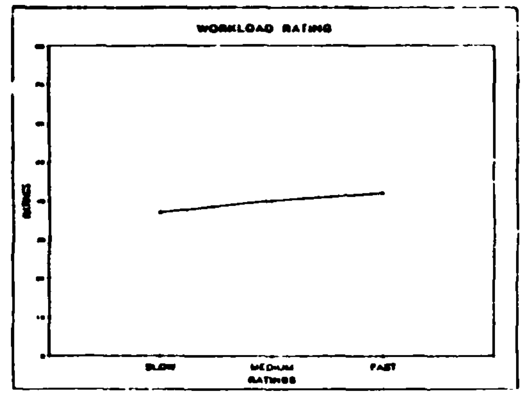
- BEHAVIOR
  - BASIC SKILLS
  - TRAINING
  - STRATEGIES
  - EFFORT EXERTED
- EMOTIONAL RESPONSE
  - MOTIVATION
  - FRUSTRATION
  - EMOTIONAL STRESS
- PHYSIOLOGICAL RESPONSE
  - FATIGUE
  - AROUSAL
  - NEURAL ACTIVITY

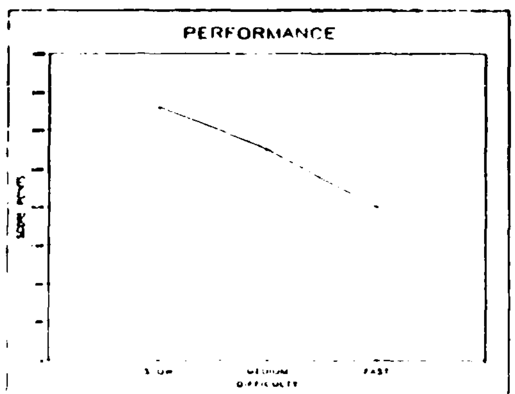


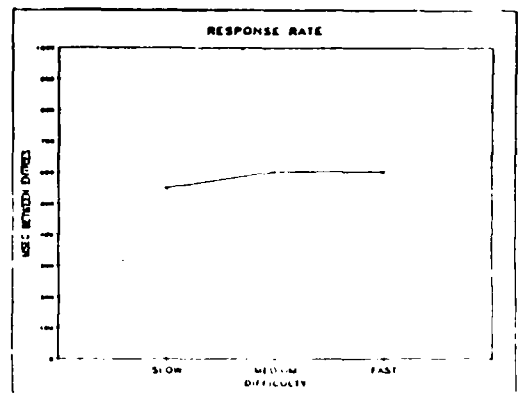
SUBJECTIVE RATINGS: INFLUENCE OF BEHAVIOR AND PERFORMANCE ON WORKLOAD RATINGS

ALTHOUGH RATINGS OFTEN COVARY WITH IMPOSED TASK DEMANDS AND PERFORMANCE, THEY MAY NOT IF OPERATORS DO NOT EXERT ADDITIONAL EFFORT AS TASK DEMANDS INCREASE. IN THIS CASE, RATINGS MAY REFLECT EFFORT RATHER THAN A PRIORI TASK DEMANDS OR ACHIEVED PERFORMANCE. THIS WAS FOUND FOR NASA-TLX RATINGS OBTAINED DURING A SUPERVISORY CONTROL SIMULATION (POPCORN).









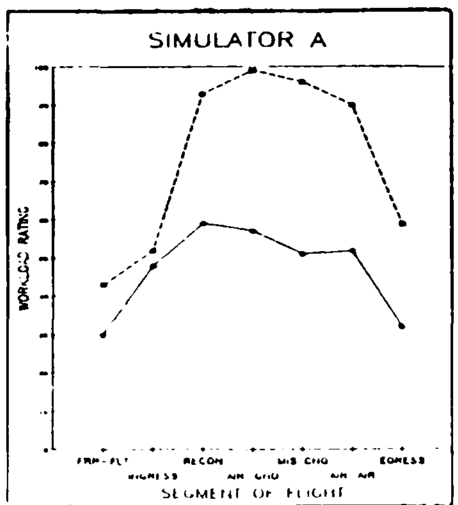
Hart & Battiste, 1984

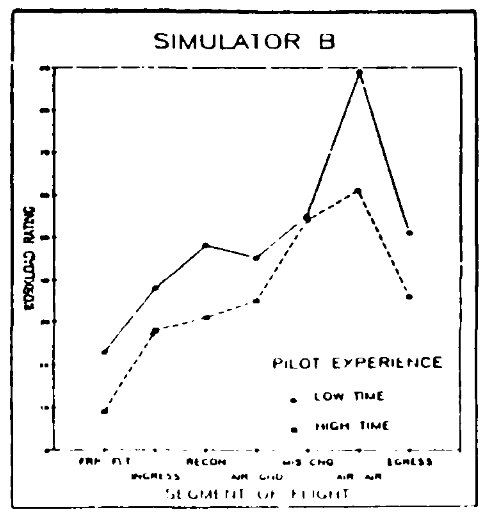
SUBJECTIVE RATINGS: BEHAVIOR

TOTAL FLIGHT HOURS, IN ADDITION TO EXPERIENCE WITH A PARTICULAR VEHICLE, INFLUENCE SUBJECTIVE RATINGS

EXAMPLE: NASA BIPOLAR RATINGS WERE OBTAINED FOR A VARIETY OF NOE FLIGHT TASKS PERFORMED IN ADVANCED HELICOPTER SIMULATORS THAT PROVIDED DIFFERENT FORMS OF AUTOMATION AND PILOT/VEHICLE INTERFACES

RESULTS: WHEN PILOTS HAD LITTLE FREEDOM TO DETERMINE WHICH TASKS TO PERFORM AND WHEN (SIMULATOR B) LOW-TIME PILOTS EXPERIENCED HIGHER WORKLOAD. WHEN THEY COULD CHOOSE WHICH TASKS TO DO AND WHEN (SIMULATOR A), HIGH-TIME PILOTS DID ADDITIONAL MISSION MANAGEMENT TASKS (COMMUNICATIONS, ASSIGNING FIRING POSITIONS), WHILE LOW-TIME PILOTS PERFORMED THE MINIMUM TASKS NECESSARY TO COMPLETE THE MISSION, THEREBY EXPERIENCING LOWER WORKLOAD.

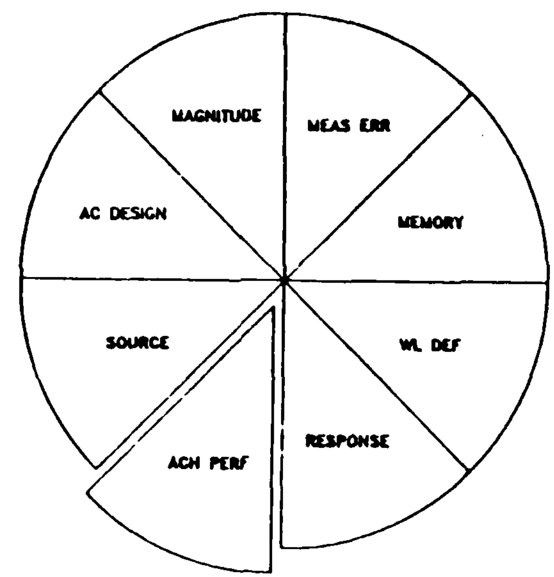




Shively, et al (in press)

SUBJECTIVE RATINGS: SOURCES OF VARIABILITY ACHIEVED PERFORMANCE

- MAY NOT REFLECT EFFORT
- INFLUENCED BY SYSTEM CHARACTERISTICS
- ADEQUACY OF PERFORMANCE FEEDBACK
- PAST PERFORMANCE AFFECTS SUBSEQUENT BEHAVIOR
- TYPE(S) OF MEASURES:
  - ACCURACY
  - SPEED
  - MULTIPLE
- EXPECTED vs ACHIEVED

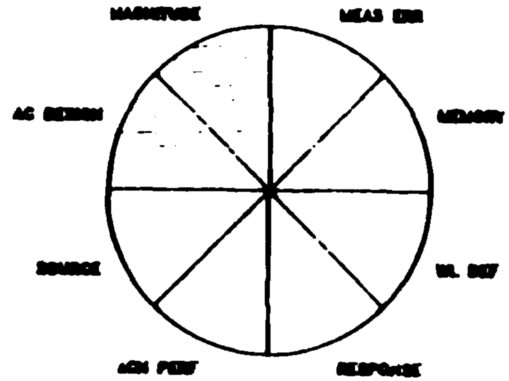


SUBJECTIVE RATINGS: EVALUATION CRITERIA

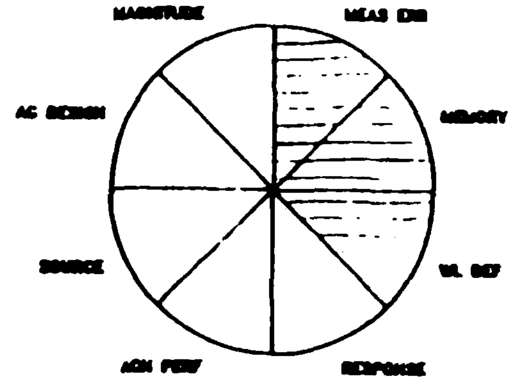
TO BE USEFUL IN AIRCRAFT CERTIFICATION, A SUBJECTIVE RATING TECHNIQUE MUST DEMONSTRATE:

- o RELIABILITY
- o SENSITIVITY
- o VALIDITY
- o DIAGNOSTICITY
- o PRACTICAL UTILITY

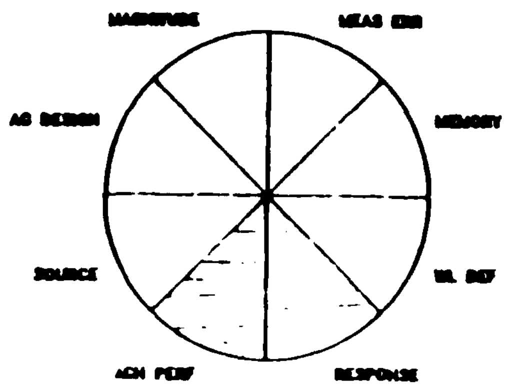
SOURCES OF RATING VARIABILITY:
FOCUS OF DIFFERENT EVALUATION PROCEDURES



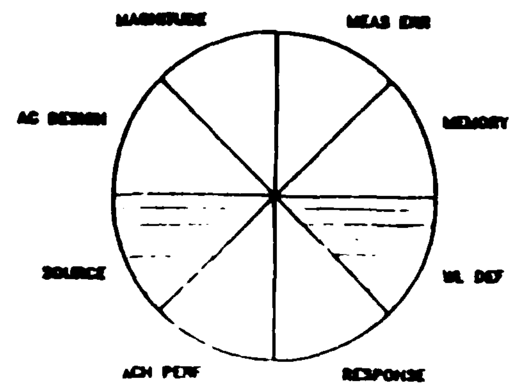
SENSITIVITY/VALIDITY



RELIABILITY



TASK X SUBJECT INTERACTION



DIAGNOSTICITY

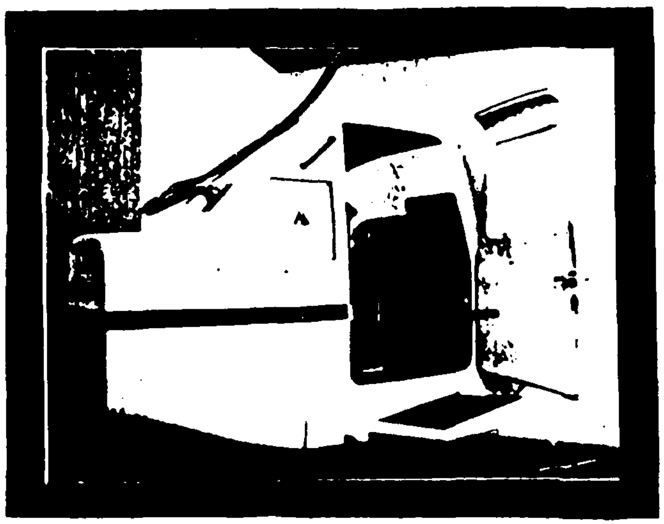
SUBJECTIVE RATINGS: RELIABILITY

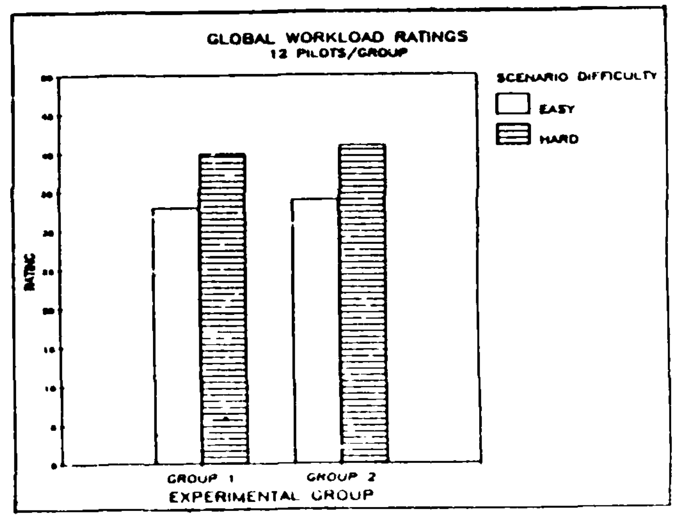
- PARTICULARLY FOR AIRCRAFT CERTIFICATION, RELIABILITY IS AN IMPORTANT CONSIDERATION BECAUSE:
  - o THE NUMBER OF EVALUATION PILOTS MAY BE LIMITED
  - o REPEATED MEASUREMENTS FOR THE SAME PILOT ARE COSTLY
  - o THE CONSEQUENCES OF AMIBGUOUS OR INACCURATE RESULTS ARE UNACCEPTABLE

- METHODS OF EVALUATING RELIABILITY:
  - o SPLIT-HALF
  - o TEST-RETEST
  - o INTER-RATER
  - o ALTERNATE FORMS

SUBJECTIVE RATING: SPLIT-HALF RELIABILITY

EXAMPLE: RATINGS WERE OBTAINED FROM TWO DIFFERENT GROUPS OF 12 PILOTS AFTER PERFORMING AN 'EASY' AND A 'HARD' FLIGHT IN A MOTION-BASE SIMULATOR USING THE NASA BIPOLAR RATING SCALE.



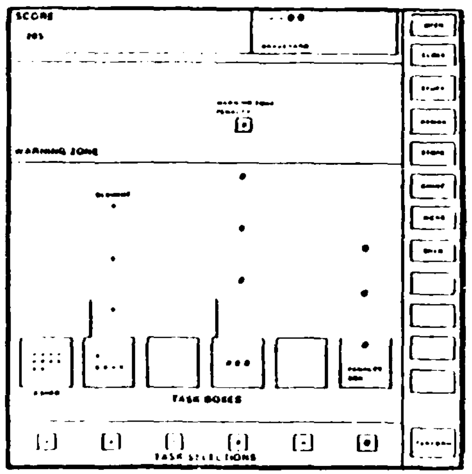


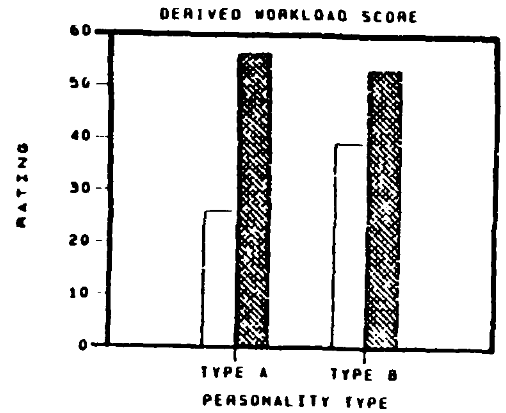
SUBJECTIVE RATINGS: SPLIT-HALF RELIABILITY

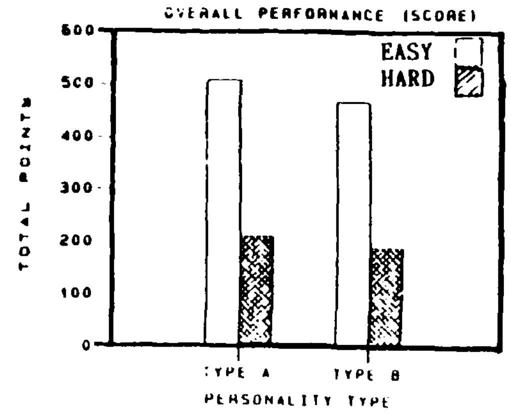
EXAMPLE: NASA BIPOLAR RATINGS WERE OBTAINED FROM SUBJECTS AFTER DIFFERENT VERSIONS OF A SUPERVISORY CONTHOL SIMULATION (POPCORN).

THEY WERE DIVIDED INTO TWO GROUPS BASED ON THEIR PERSONALITY TYPE.

WORKLOAD RATINGS AND PERFORMANCE WERE NOT SIGNIFICANTLY DIFFERENT BETWEEN THE GROUPS (ALTHOUGH BLOOD PRESSURE RESPONSIVENESS WAS).







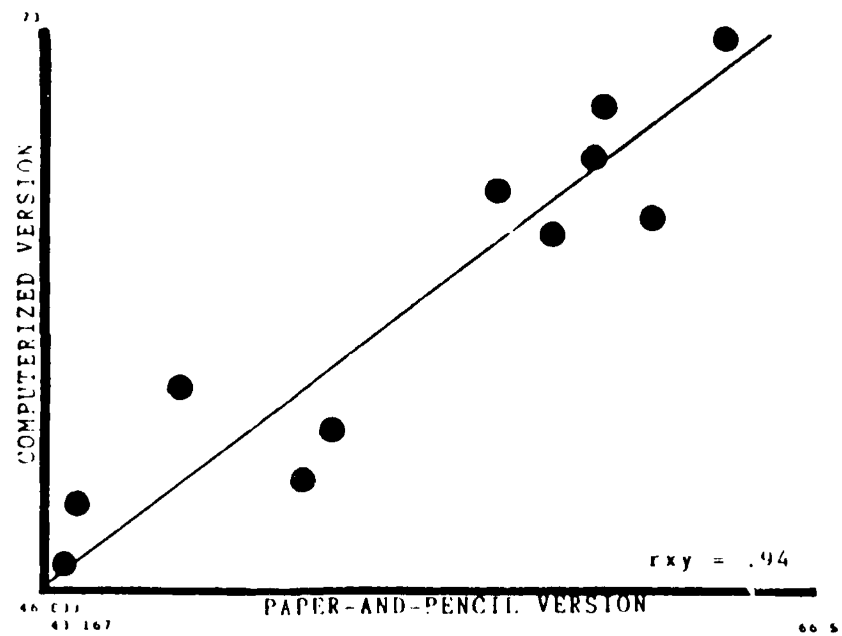
Hart et al, in press

SUBJECTIVE RATINGS: ALTERNATE FORMS RELIABILITY

EXAMPLE: RATINGS WERE OBTAINED FROM SIX SUBJECTS USING THREE FORMS OF THE NASA-TASK LOAD INDEX: COMPUTERIZED, PAPER-AND-PENCIL, AND VERBAL

TASKS: TARGET ACQUISITION, GRAMMATICAL REASONING, UNSTABLE TRACKING

RESULTS: RATINGS WITH ALTERNATE FORMS OF NASA-TLX WERE HIGHLY CORRELATED



King, 1987

SUBJECTIVE RATINGS: GENERAL APPROACHES TO VALIDATION

- WORKLOAD MEASURES MUST BE EVALUATED AGAINST CRITERIA THAT THEY CAN BE LOGICALLY EXPECTED TO FULFILL AND THAT ARE PRACTICALLY RELEVANT TO A GIVEN RESEARCH QUESTION
- ALTHOUGH SUBJECTIVE RATINGS HAVE A 'COMPELLING SENSE OF RELEVANCE BECAUSE THEY DEPEND DIRECTLY ON THE RATER'S PERSONAL EXPERIENCES' (GOPHER & DONCHIN, 1986)
- THEY ARE DIFFICULT TO VALIDATE BECAUSE THEY ARE DESIGNED TO MEASURE AN INDIVIDUAL'S PERSONAL ASSESSMENT OF AN EXPERIENCE RATHER THAN OBJECTIVE CHARACTERISTICS OF THE TASK, AIRCRAFT DESIGN, THE ENVIRONMENT, OR SYSTEM PERFORMANCE

SUBJECTIVE RATINGS: CONSTRUCT VALIDITY APPROACHES

THE TYPICAL SOLUTION TO THE PROBLEM OF PROVIDING INDEPENDENT VALIDATION FOR SUBJECTIVE RATING SCALES IS TO COMPARE RATINGS AGAINST OTHER, MORE "OBJECTIVE" AND MEASURABLE CRITERIA THAN THE PERSONAL EXPERIENCE OF THE RATER:

* THEORY-BASED PREDICTIONS OF TASK DEMANDS
* MEASURES OF PERFORMANCE

. . . HOWEVER, SUBJECTIVE RATINGS DO NOT REFLECT THESE ASPECTS OF WORKLOAD DIRECTLY

AN ALTERNATIVE SOLUTION IS TO OBTAIN:

* CONVERGING EVIDENCE FROM OTHER MEASURES
* AN ACCUMULATION OF EVIDENCE FROM REPEATED USE

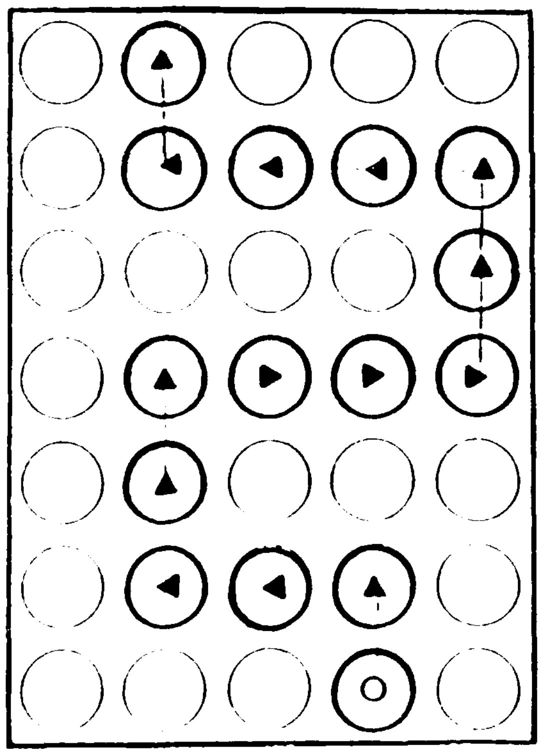
SUBJECTIVE RATINGS: CONSTRUCT VALIDITY ACCUMULATION OF EVIDENCE

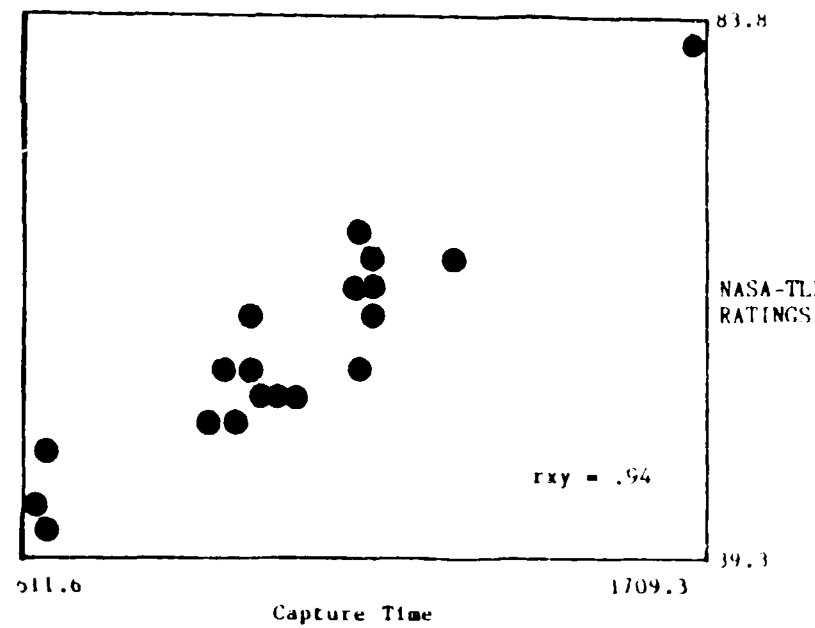
THE MOST COMMON METHOD OF VALIDATING A CANDIDATE SCALE IS TO OBTAIN AN ACCUMULATION OF EVIDENCE THAT IT IS SENSITIVE TO SUBTLE AS WELL AS GROSS VARIATIONS IN:

- THEORY-BASED ESTIMATES OR EXPERT GUESSES ABOUT TASK DEMANDS
- THEORY-BASED ESTIMATES OR EXPERT GUESSES ABOUT THE EFFECTS OF DISPLAYS, CONTROLS, AND AUTOMATED SUBSYSTEMS
- MEASURABLE LEVELS OF EFFORT EXERTED IN PERFORMING TASKS
- ACHIEVED LEVELS OF PERFORMANCE
- DIFFERENTIAL SENSITIVITY TO APPARENT VARIATIONS IN THE SOURCES OF TASK DEMANDS
- ADDITIONAL INFORMATION PROVIDED BY THE RATERS OR OBSERVERS

SUBJECTIVE RATING: CONSTRUCT VALIDITY

EXAMPLE: THE RELATIONSHIP BETWEEN PERFORMANCE (CAPTURE TIME) AND NASA TASK LOAD INDEX RATINGS WAS EVALUATED FOLLOWING A SERIES OF SEQUENTIAL TARGET ACQUISITIONS (TIMEPOOLS). DIFFICULTY WAS MANIPULATED BY VARYING TARGET SIZE, RATE OF CHANGE OF TARGET SIZE, AND RESPONSE SELECTION LOAD.



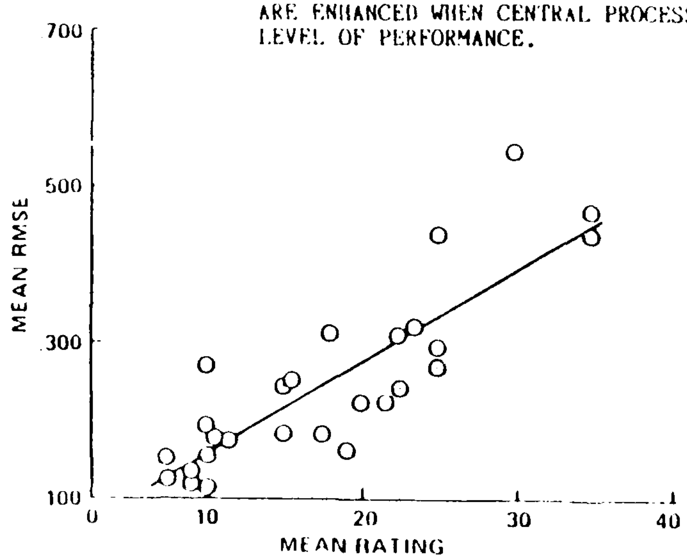


Johnson (in press)

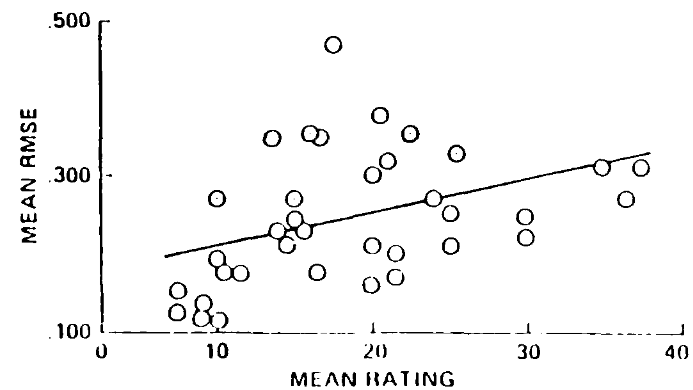
SUBJECTIVE RATING: CONSTRUCT VALIDITY

EXAMPLE: THE RELATIONSHIP BETWEEN PERFORMANCE (RMS TRACKING ERROR) AND GLOBAL WORKLOAD RATINGS WAS EVALUATED FOLLOWING INTERVALS FILLED WITH A SINGLE-AXIS COMPENSATORY TRACKING TASK. DIFFICULTY WAS VARIED BY MANIPULATING FORCING FUNCTION BANDWIDTH AND ORDER OF CONTROL.

RATING X PERFORMANCE CORRELATIONS WERE HIGHER WITH ORDER-OF-CONTROL MANIPULATIONS; ORDER OF CONTROL PRIMARILY INFLUENCES THE CENTRAL PROCESSING DEMANDS OF A TASK, WHEREAS BANDWIDTH PRIMARILY INFLUENCES RESPONSE EXECUTION PROCESSING. THUS, RATING X PERFORMANCE CORRELATIONS ARE ENHANCED WHEN CENTRAL PROCESSING IS THE LIMITING FACTOR FOR THE LEVEL OF PERFORMANCE.



(a) ORDER MANIPULATION



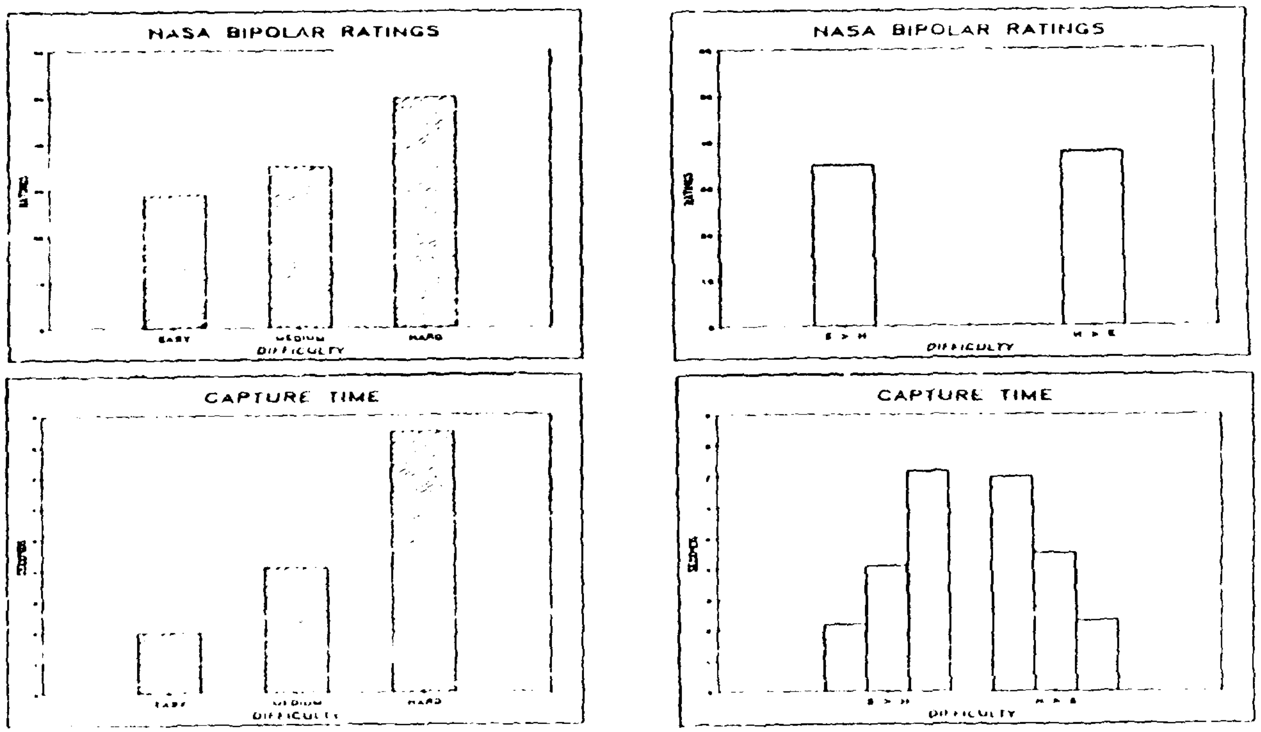
(b) BANDWIDTH MANIPULATION

Vidulich, in press

SUBJECTIVE RATINGS: CONSTRUCT VALIDITY

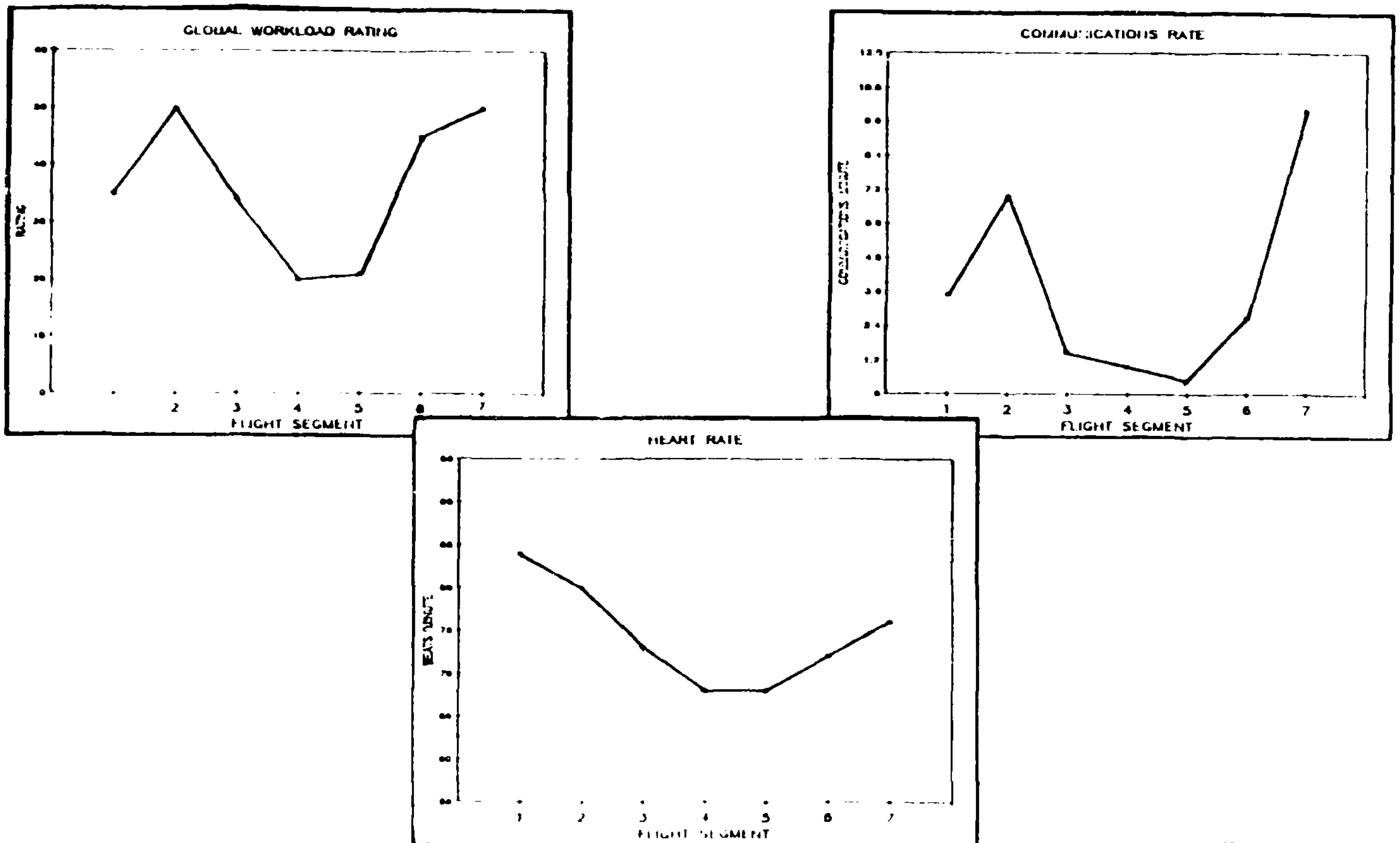
RATING SCALES MUST INTEGRATE MOMENTARY WORKLOAD DEMANDS ACROSS TIME, GIVING EQUAL WEIGHTS TO EVENTS THAT HAPPENED EARLY AND LATE IN THE INTERVAL.

EXAMPLE: NASA BIPOLAR RATINGS REFLECTED THE AVERAGE WORKLOAD OF INTERVALS FILLED WITH TASKS THAT INCREASED, DECREASED, OR REMAINED CONSTANT IN DIFFICULTY WITHIN THE INTERVAL TO BE ESTIMATED.



SUBJECTIVE RATING: CONSTRUCT VALIDITY

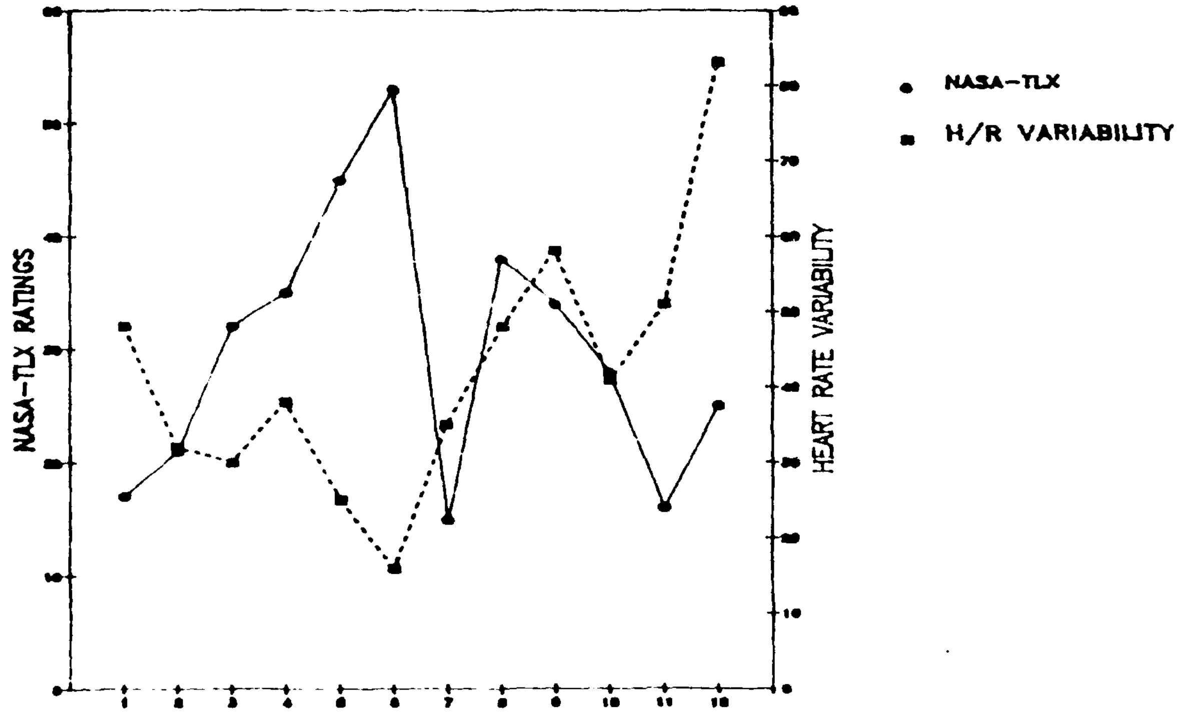
EXAMPLE: CONVERGING EVIDENCE WAS PROVIDED BY MULTIPLE MEASURES OF WORKLOAD TAKEN DURING 11 FLIGHTS OF THE NASA C-141 KUIPER AIRBORNE OBSERVATORY.



Hart & Hauser, in press

SUBJECTIVE RATING: CONSTRUCT VALIDITY

EXAMPLE: COMPARISON BETWEEN NASA-TASK LOAD INDEX RATINGS AND HEART RATE VARIABLITY MEASURES OBTAINED DURING EXPERIMENTAL FLIGHTS IN THE NASA SH-3G HELICOPTER.

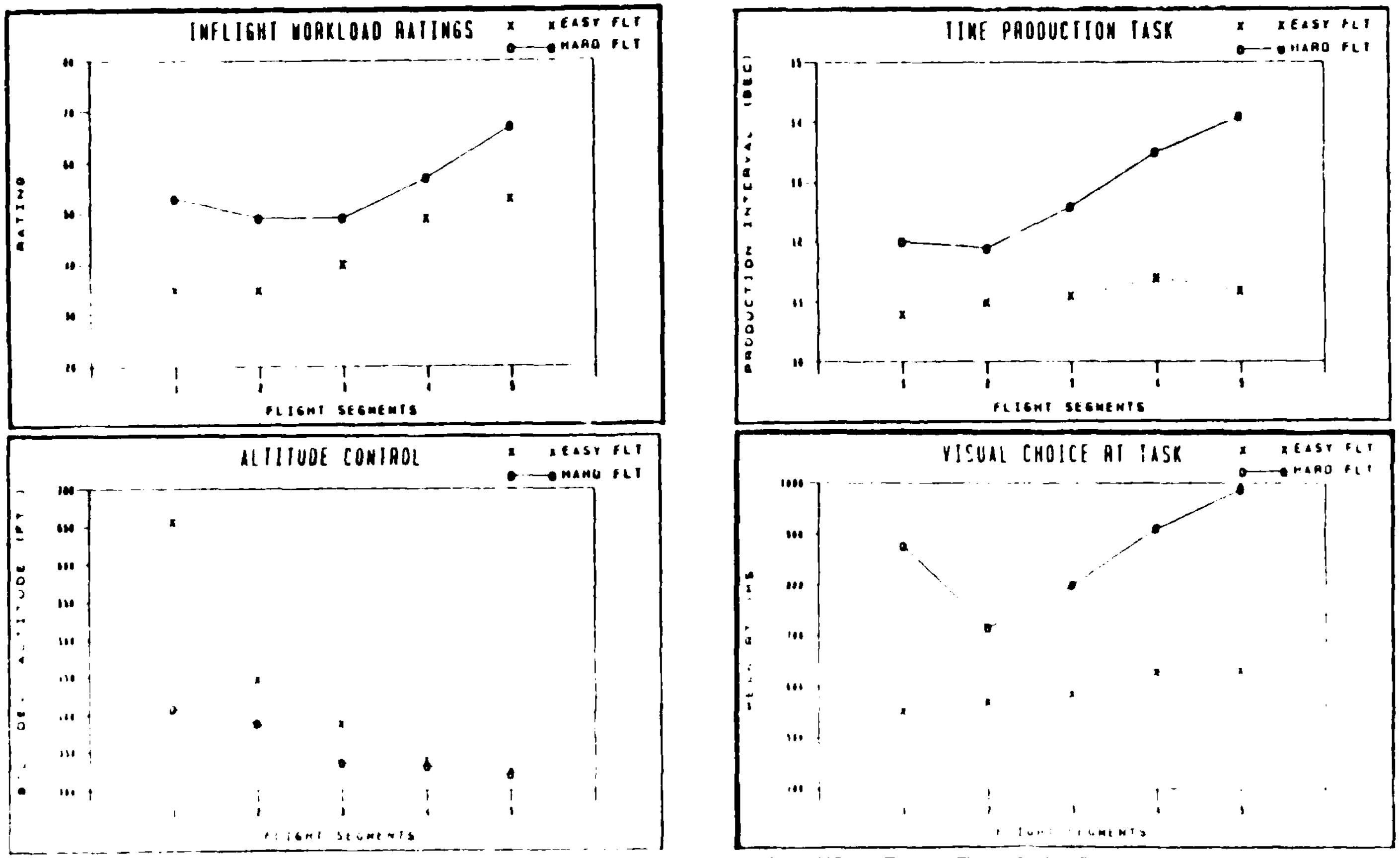


NOTE: A NEGATIVE CORRELATION WAS PREDICTED AND FOUND (-.43)

Shively et al., in pre...

SUBJECTIVE RATING: CONSTRUCT VALIDITY

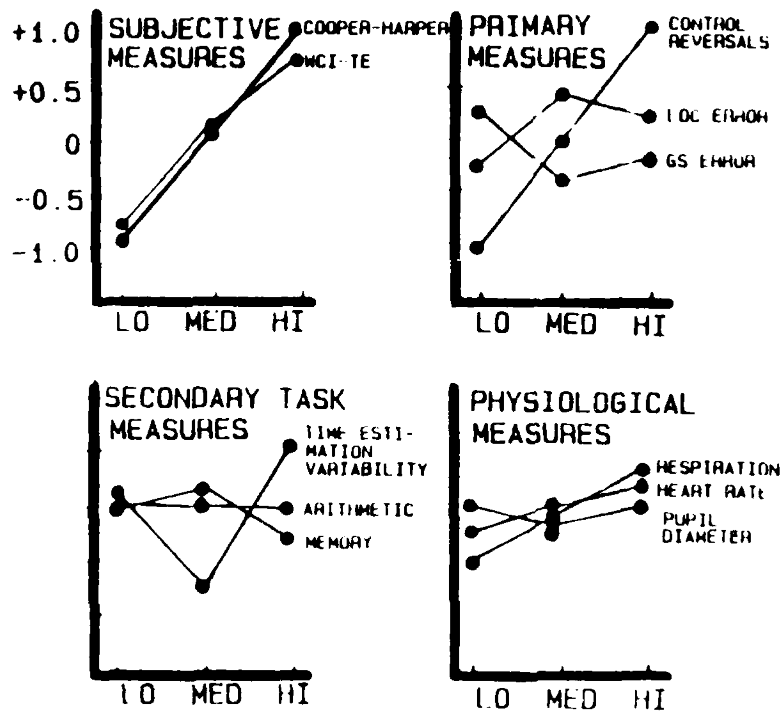
EXAMPLE: CONVERGING EVIDENCE WAS PROVIDED BY MULTIPLE MEASURES OF WORKLOAD OBTAINED DURING INSTRUMENT FLIGHTS PERFORMED IN A MOTION-BASE SIMULATOR. FLIGHT-SEGMENT DIFFICULTY WAS ESTABLISHED A PRIORI.

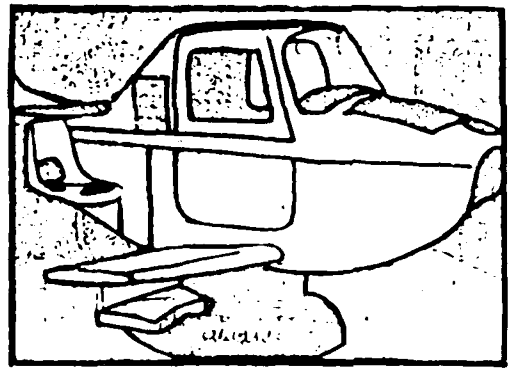


Bortolussi et al, 198

SUBJECTIVE RATING: CONSTRUCT VALIDITY

EXAMPLE: LACK OF CONVERGING EVIDENCE AMONG MANY TYPES OF WORKLOAD MEASURES OBTAINED DURING A SIMULATION EXPERIMENT THAT VARIED LEVELS OF PERCEPTUAL/MOTOR LOAD. WHICH EVIDENCE DO YOU BELIEVE? WERE THE WORKLOAD LEVELS REALLY DIFFERENT?





ADDITIONAL INSENSITIVE MEASURES:
DIGIT SHADOWING, EYEBLINK FREQUENCY, VOICE STRESS.

Wierwille & Connor, 1983

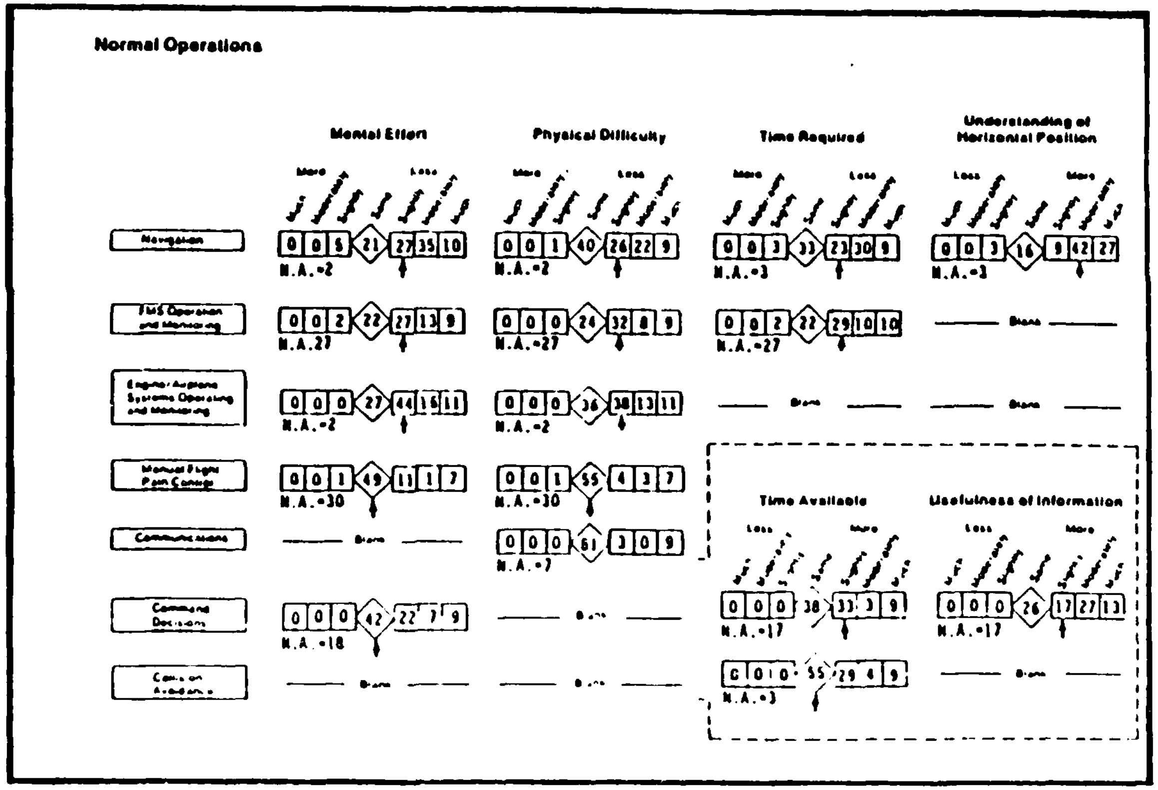
SUBJECTIVE RATINGS: DIAGNOSTICITY

TO BE USEFUL IN AIRCRAFT CERTIFICATION, A MEASURE MUST PROVIDE DIAGNOSTIC INFORMATION ABOUT THE SPECIFIC SOURCES OF WORKLOAD THAT WERE RELEVANT IN A PARTICULAR TASK

- PSYCHOLOGICAL VARIABLES
  - PHYSICAL DEMANDS/EFFORT
  - MENTAL DEMANDS/EFFORT
  - STRESS
  - TIME PRESSURE
- TASK-RELATED VARIABLES
  - FLIGHT PATH CONTROL
  - COLLISION AVOIDANCE
  - NAVIGATION
  - COMMUNICATIONS
  - COMMAND DECISIONS
- ENVIRONMENTAL VARIABLES
  - SOCIAL
  - PHYSICAL

SUBJECTIVE RATINGS: DIAGNOSTICITY

EXAMPLE: DISTRIBUTIONS OF RATINGS OBTAINED DURING CERTIFICATION FLIGHTS FOR THE B-767 DURING DEPARTURE AND ARRIVAL USING THE PILOT SUBJECTIVE EVALUATION SCALE



Ruggiero & Fadden, 1982

SUBJECTIVE RATINGS: REFERENCE, BENCHMARK, OR ANCHOR TASKS

- REFERENCE TASKS CAN:
  - CALIBRATE RATERS
  - REDUCE INDIVIDUAL DIFFERENCES
  - STANDARDIZE THE FACTORS CONSIDERED IN PROVIDING RATINGS
  - HELP IN COMMUNICATING RESULTS

- OPERATIONALLY, THEY CAN BE EITHER:
  - PERFORMED AND RATED FOR COMPARISON WITH TARGET TASK
  - PERFORMED AND POSITIONED ARBITRARILY ON A SCALE AS AN ANCHOR - RATINGS ARE MADE RELATIVE TO THAT POSITION.
  - DESCRIBED, RATED FROM PILOTS' RECOLLECTIONS OF SIMILAR EXPERIENCES,AND COMPARED TO TARGET TASK.

SUBJECTIVE RATINGS: USE OF A REFERENCE TASK (1)

EXAMPLE: MAGNITUDE ESTIMATES OF TASK DIFFICULTY WERE OBTAINED (RELATIVE TO A SINGLE-AXIS TRACKING TASK REFERENCE) FOR DIFFERENT SINGLE-AND DUAL-TASK CONDITIONS.

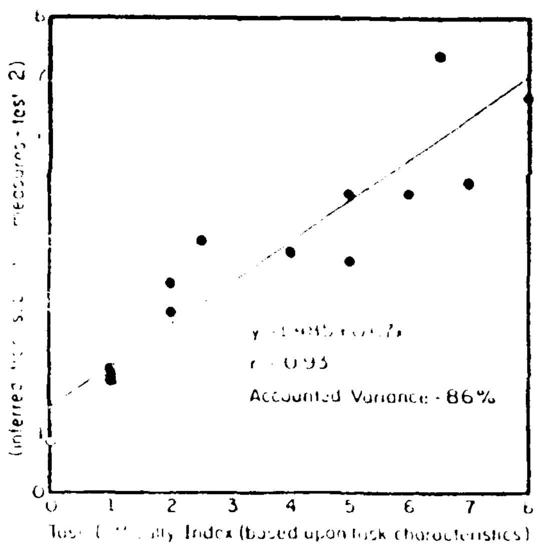


Fig. 11 - The relationship between the index of task difficulty and units of invested resources derived from subjective estimates.

Split Half Reliability for the Average Scores of the 21 Conditions

Raw Scores		Transformed	
Practice	.98	Practice	.98
Test 1	.96	Test 1	.98
Reevaluation 1	.97	Reevaluation 1	.98
Test 2	.97	Test 2	.98
Reevaluation 2	.97	Reevaluation 2	.97

(N first half = 29; N second half = 26)

Gopher & Braune, 1984

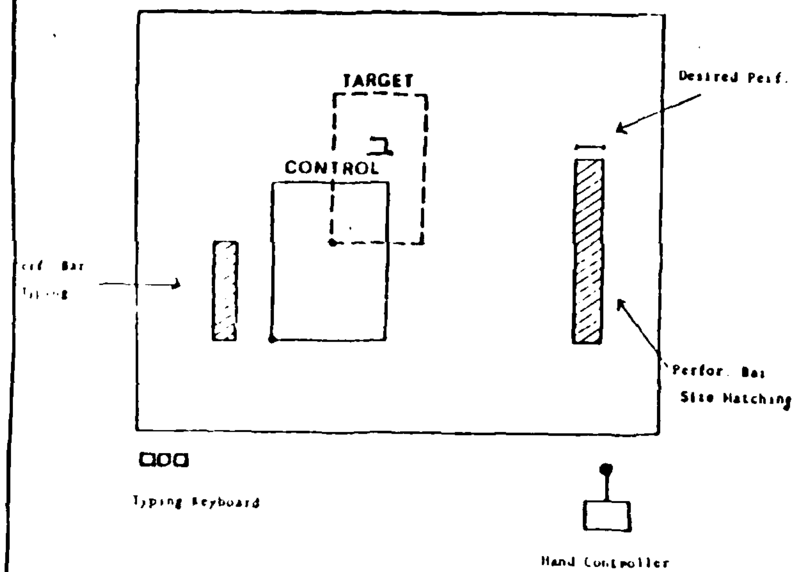
SUBJECTIVE RATINGS: USE OF A REFERENCE TASK (2)

EXAMPLE: MAGNITUDE ESTIMATES OF TASK DIFFICULTY WERE OBTAINED FOR DIFFERENT SINGLE- AND DUAL-TASK COMBINATIONS RELATIVE TO:

(1) A SINGLE-AXIS TRACKING TASK

(2) A LETTER-ENTRY TASK

RESULTS: THE MAGNITUDE OF THE RATINGS WAS INFLUENCED BY THE REFERENCE TASK USED



SUBJECT DISPLAY:
DUAL TASK - SIZE MATCHING/TYPING

Reference task effect on load estimates of the letter typing task

Conditions / Reference Tasks	Ss	SD	D Primary	Overall
Reference Size Matching	82.43	81.56	138.70	100.90
Reference Letter Typing	98.98	96.20	116.80	103.99

Reference task effect on load estimates of the size matching task

Conditions / Reference Tasks	Ss	SD	D Primary	Overall
Reference Size Matching	110.47	101.57	139.87	117.30
Reference Letter typing	95.14	91.89	119.07	102.03

Ss - Single Task in Single Session
SD - Single Task in Dual Session
D Primary - Primary Task in Dual Session

Gopher, Chillag & Arzi, 1985

SUBJECTIVE RATINGS: PRACTICAL ISSUES

- TO BE PRACTICALLY USEFUL, RESPONSES TO A RATING SCALE SHOULD BE EASY TO OBTAIN AND SHOULD NOT INTERFER WITH TASK PERFORMANCE

- IF TASK PERFORMANCE IS DEGRADED OR ALTERED BY ADMINISTERING THE MEASURE, THE VALIDITY OF THE INFORMATION IT PROVIDES IS SUSPECT FOR:
    - PRACTICAL REASONS
    - THEORETICAL REASONS

SUBJECTIVE RATINGS: AVAILABLE SCALES

- UNIDIMENSIONAL SCALES
  - 10-CM LINE
  - McDONNELL
  - MAGNITUDE ESTIMATION
  - PAIRWISE COMPARISONS
- DECISION-TREE FORMAT
  - COOPER-HARPER HANDLING QUALITIES RATING SCALE (HQR)
  - WOLFE
  - SIMPSON-SHERIDAN
  - BEDFORD
  - MODIFIED COOPER-HARPER (MCH)
- MULTIDIMENSIONAL
  - SIMPSON-SHERIDAN
  - WORKLOAD/COMPENSATION INTERFERENCE/ TECHNICAL EFFECTIVENESS (WCI/TE)
  - SUBJECTIVE WORKLOAD ASSESSMENT TECHNIQUE (SWAT)
  - NASA BIPOLAR RATING SCALE
  - NASA-TASK LOAD INDEX (TLX)
  - BOEING RATING SCALE

UNIDIMENSIONAL WORKLOAD RATING SCALES

SUBJECTIVE RATINGS: McDONNELL SCALE

DEMANDS ON PILOT

0

1

2

- Completely undemanding, very relaxed and comfortable

3

- Largely undemanding, relaxed

4

5

- Mildly demanding of pilot attention, skill, or effort

6

- Demanding of pilot attention, skill, or effort

7

- Very demanding of pilot attention, skill, or effort

8

- Completely demanding of pilot attention, skill, or effort

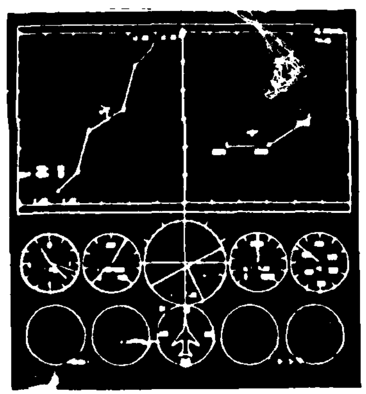
9 - Nearly uncontrollable

10 Uncontrollable

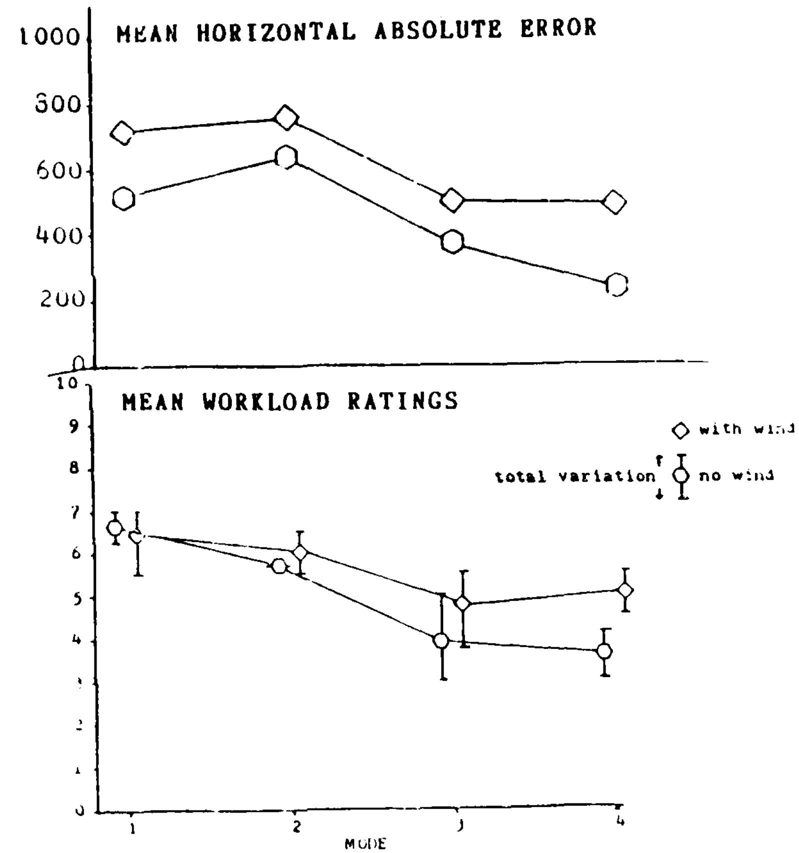
Not applicable

SUBJECTIVE RATINGS: McDONNELL SCALE

EXAMPLE: COMPARISON OF McDONNELL SCALE RATINGS AND PERFORMANCE IN A SIMULATED YC-15. DIFFICULTY WAS MANIPULATED BY VARYING WIND GUSTS ACROSS SUCCESSIVELY HIGHER ORDERS OF CONTROL AUTOMATION (MODES)



Computer Generated Display

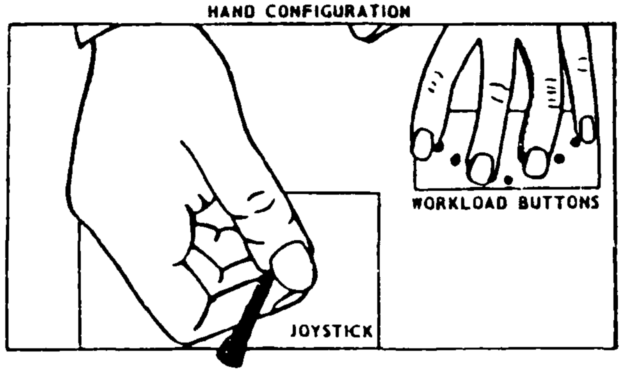


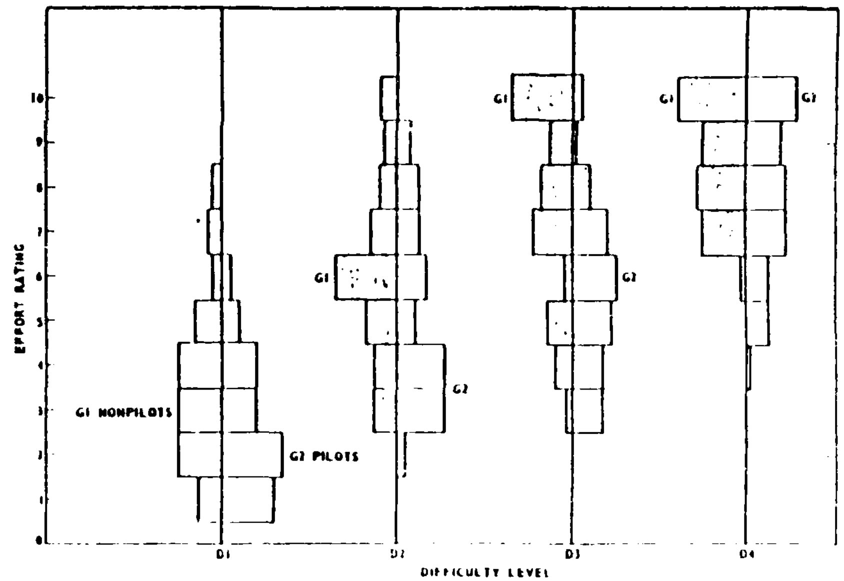
Sheridan, Yoerger, Tulga, & Daryanian, 1978

RATING SCALES: UNIDIMENSIONAL SCALE
PILOT OBJECTIVE/SUBJECTIVE WORKLOAD ASSESSMENT TECHNIQUE (POSWAT)

EXAMPLE: DISCRETE DIFFICULTY RATINGS (10 POSSIBLE VALUES) WERE OBTAINED AT 1-MIN INTERVALS DURING FOUR LEVELS OF A JEX UNSTABLE TRACKING TASK

DETAIL OF SUBJECT'S HAND CONFIGURATION





Rosenberg, Stein, & Rehmann, 1981

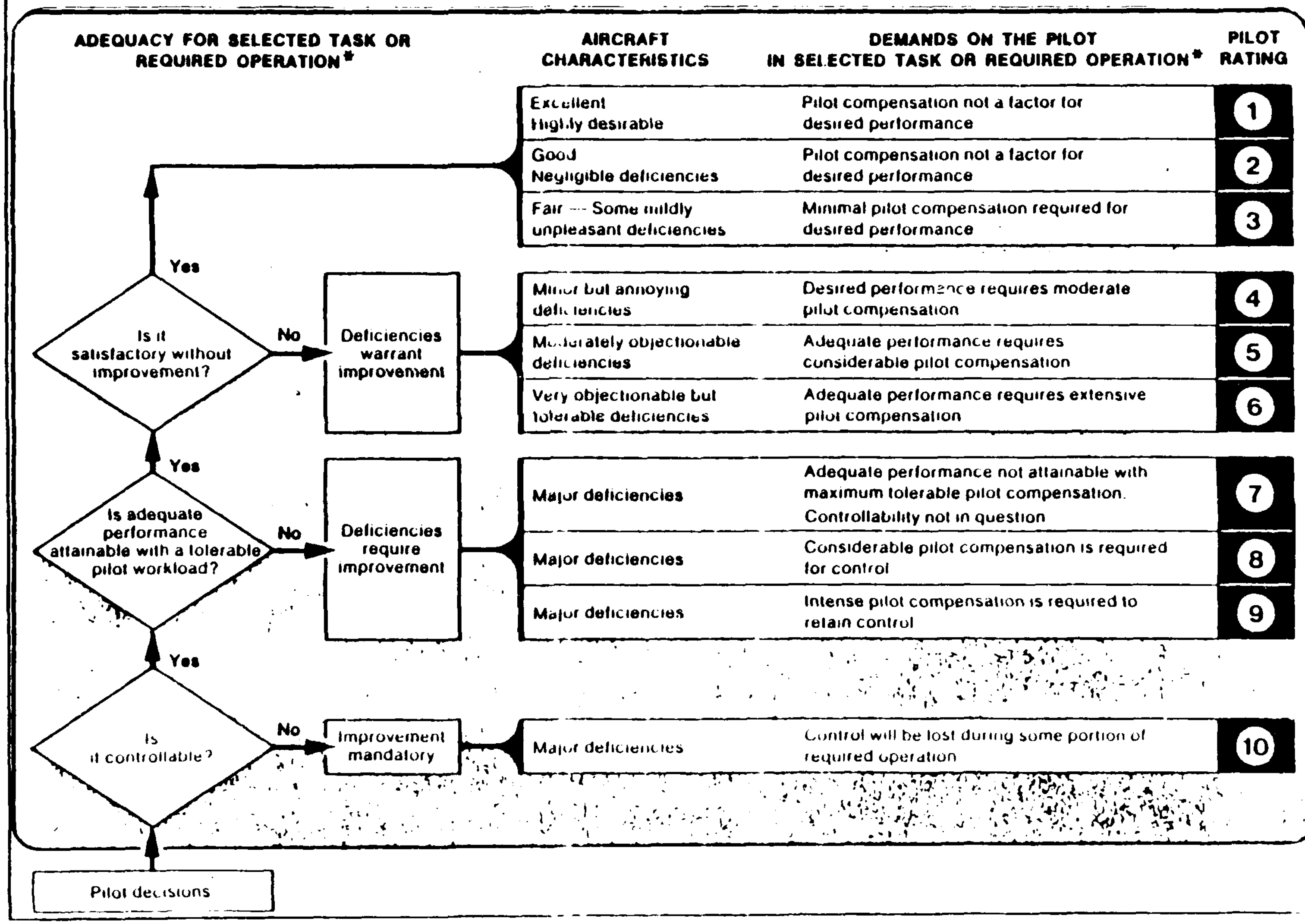
DECISION-TREE FORMAT SCALES

SUBJECTIVE RATINGS: ORIGINAL COOPER AND HARPER SCALES

COOPER					PR
[illegible]	SUBJECTIVE RATING	MISSION	PRIMARY MISSION ACCOMPLISHED?	CAN BE LANDED	
[illegible]			Yes	Yes	1
Good, [illegible]	Satisfactory	Normal operation	Yes	Yes	2
[illegible]			Yes	Yes	3
[illegible]			Yes	Yes	4
[illegible] operation	Unsatisfactory	Emergency operation	Doubtful	Yes	5
[illegible]			Doubtful	Yes	6
[illegible]			No	Doubtful	7
[illegible]	Unacceptable	No operation	No	No	8
[illegible]			No	No	9
[illegible]	Unprintable	What [illegible]			10

[illegible]		PILOT ATTENTION OR EFFORT REQUIRED	ADJECTIVAL DESCRIPTION WITHIN CATEGORY	PR
FLYING QUALITIES	AIRCRAFT ACCEPTABILITY			
SATISFACTORY Criterion: Pilot performance is not seriously affected by any flying quality deficiencies which may be present Criterion: [illegible] seriously affect [illegible] pilot would [illegible] that the [illegible] characteristics [illegible] be improved	ACCEPTABLE		Excellent	1
			Good	2
			Fair	3
UNSATISFACTORY [illegible] performance is [illegible] affected by flying quality deficiencies that pilot [illegible] that characteristics be [illegible]	"RELUCTANTLY" ACCEPTABLE Criterion: [illegible] performance deficiencies cannot be improved without a serious compromise of the other factors which influence the mission capability of the airplane		Fair	4
			Poor	5
			Bad	6
	UNACCEPTABLE	Required major portion of pilot's attention [illegible]	Bad	7
		[illegible]	Very bad	8
		Aircraft just [illegible]	Dangerous	9
	UNFLYABLE	[illegible]	Unflyable	10

RATING SCALES: COOPER-HARPER HANDLING QUALITIES RATING SCALE



RATING SCALES: COOPER-HARPER HANDLING QUALITIES RATING SCALE

DEFINITIONS FROM TN-D-5153

COMPENSATION

The measure of additional pilot effort and attention required to maintain a given level of performance in the face of deficient vehicle characteristics.

HANDLING QUALITIES

Those qualities or characteristics of an aircraft that govern the ease and precision with which a pilot is able to perform the tasks required in support of an aircraft role.

MISSION

The composite of pilot-vehicle functions that must be performed to fulfill operational requirements. May be specified for a role, complete flight, flight phase, or flight subphase.

PERFORMANCE

The precision of control with respect to aircraft movement that a pilot is able to achieve in performing a task. (Pilot-vehicle performance is a measure of handling performance. Pilot performance is a measure of the manner or efficiency with which a pilot moves the principal controls in performing a task.)

ROLE

The function or purpose that defines the primary use of an aircraft.

TASK

The actual work assigned a pilot to be performed in completion of or as representative of a designated flight segment.

WORKLOAD

The integrated physical and mental effort required to perform a specified piloting task.

SUBJECTIVE RATINGS: COOPER-HARPER HANDLING QUALITIES RATINGS

EXAMPLE: OBTAINED COOPER-HARPER HQS FROM A NUMBER OF EXPERIMENTS WERE COMPARED TO AN INTEGRATED INDEX OF TASK PERFORMANCE

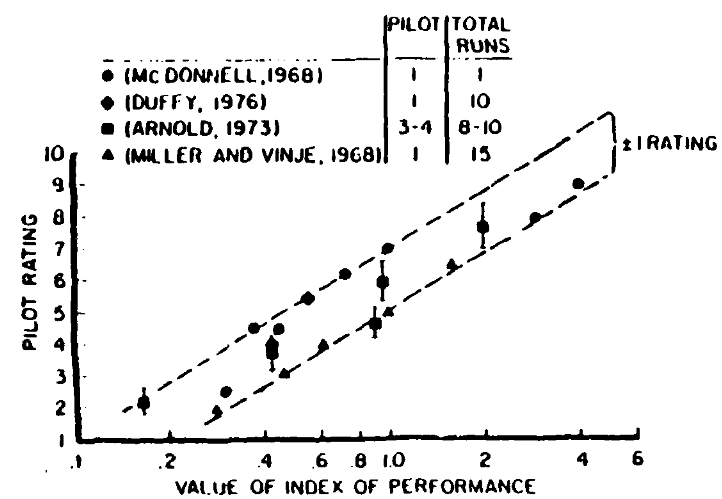
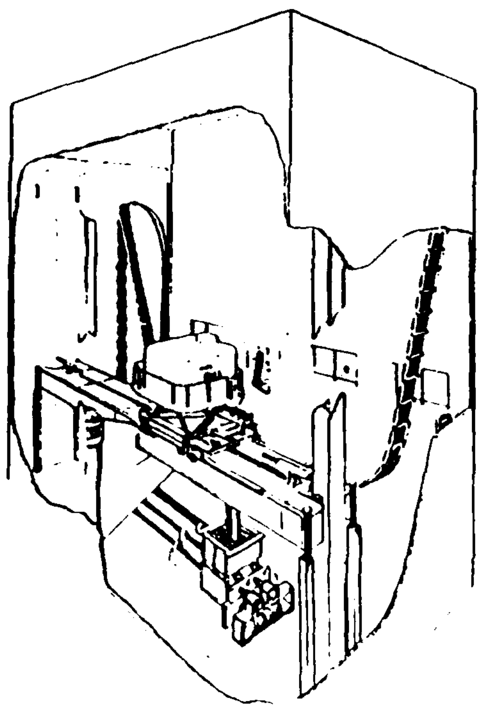


Figure 4 Pilot rating vs. value of model index of performance

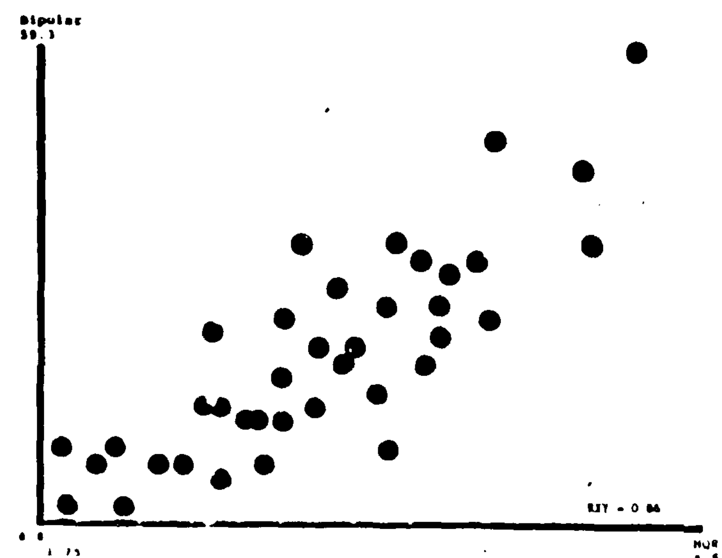
Hess, 1977

SUBJECTIVE RATINGS: COOPER-HARPER HQRs (CONVERGING EVIDENCE OF VALIDITY)

EXAMPLE: WORKLOAD RATINGS AND HQR RATINGS WERE OBTAINED DURING AN ADVANCED HELICOPTER SIMULATION. LEVEL OF AUTOMATION AND CREW SIZE (ONE vs TWO) WERE VARIED. IN THIS SITUATION, THE TWO MEASURES REFLECTED THE SAME FACTORS (E.G., WORKLOAD)



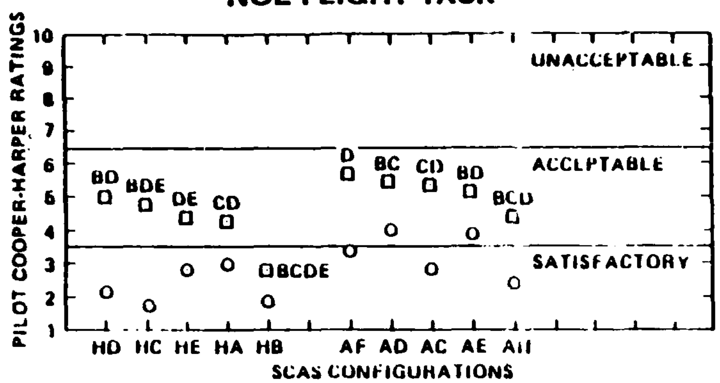
NASA BIPOLAR vs COOPER-HARPER HQRs



Haworth, Bivens, & Shively, 1986

SUBJECTIVE RATINGS: COOPER-HARPER HQRs

NOE FLIGHT TASK

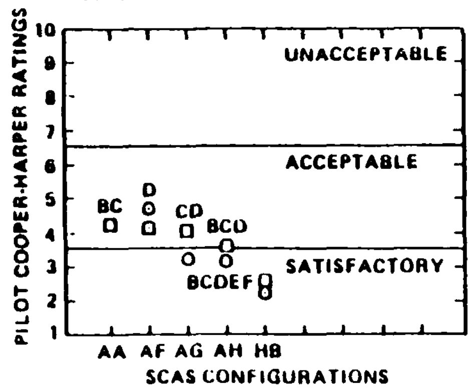


□ SINGLE PILOT
○ DUAL PILOT

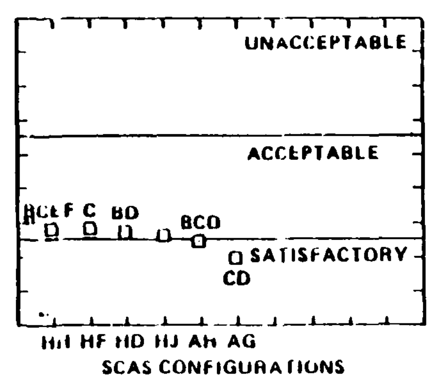
SCAS CONFIGURATIONS:
- A–ATTITUDE COMMAND/STABILIZATION
- B–ADOCS HYBRID CONTROL SYSTEM

PILOT-SELECTABLE FEATURES:
- B–TURN COORDINATION
- C–HEADING HOLD
- D–ALTITUDE HOLD
- E–AIRSPEED HOLD
- F–POSITION HOLD

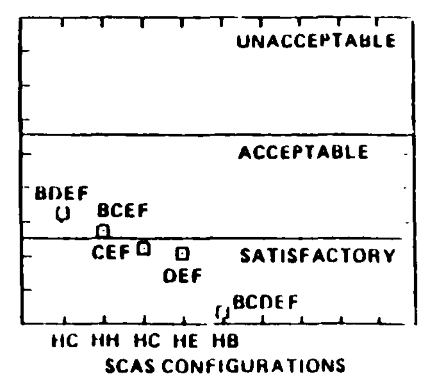
AIR-TO-GROUND ATTACK



AIR-TO-AIR ENGAGEMENT TASK

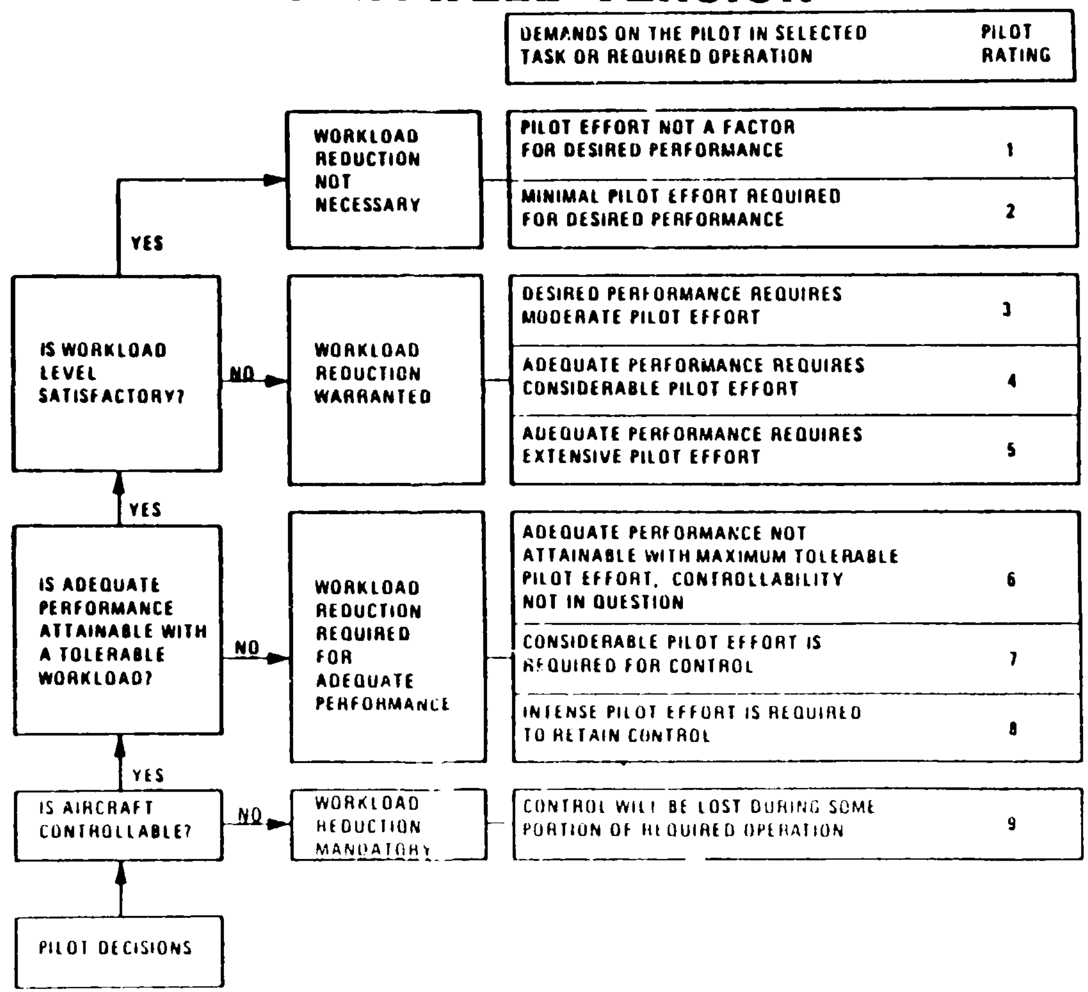


HOVER TASK



Haworth, Bivens, & Shively, 1986

RATING SCALES: MODIFIED COOPER-HARPER SCALE
HONEYWELL VERSION

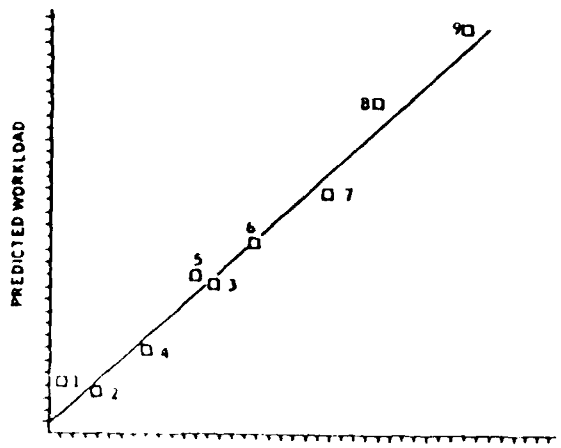


Wolfe, 1977; Stackhouse & Petersen, 1972
North, Stackhouse &Graffunder, 1979

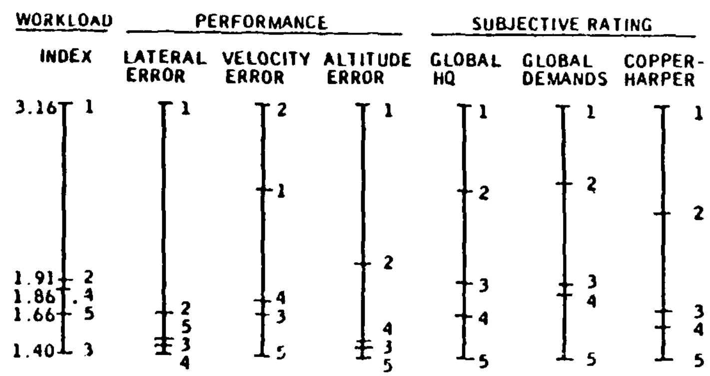
RATING SCALES: MODIFIED COOPER-HARPER SCALE HONEYWELL VERSION

EXAMPLE: RATINGS WERE COMPARED TO A WORKLOAD INDEX DERIVED FROM MULTIPLE FEATURES OF PHYSIOLOGICAL MEASURES

9 LEVELS OF TRACKING TASK DIFFICULTY (3 WIND GUST LEVELS X 3 ORDER OF CONTROL LEVELS)



5 LATERAL-AXIS-OF-CONTROL MODELS FOR AN F-4C



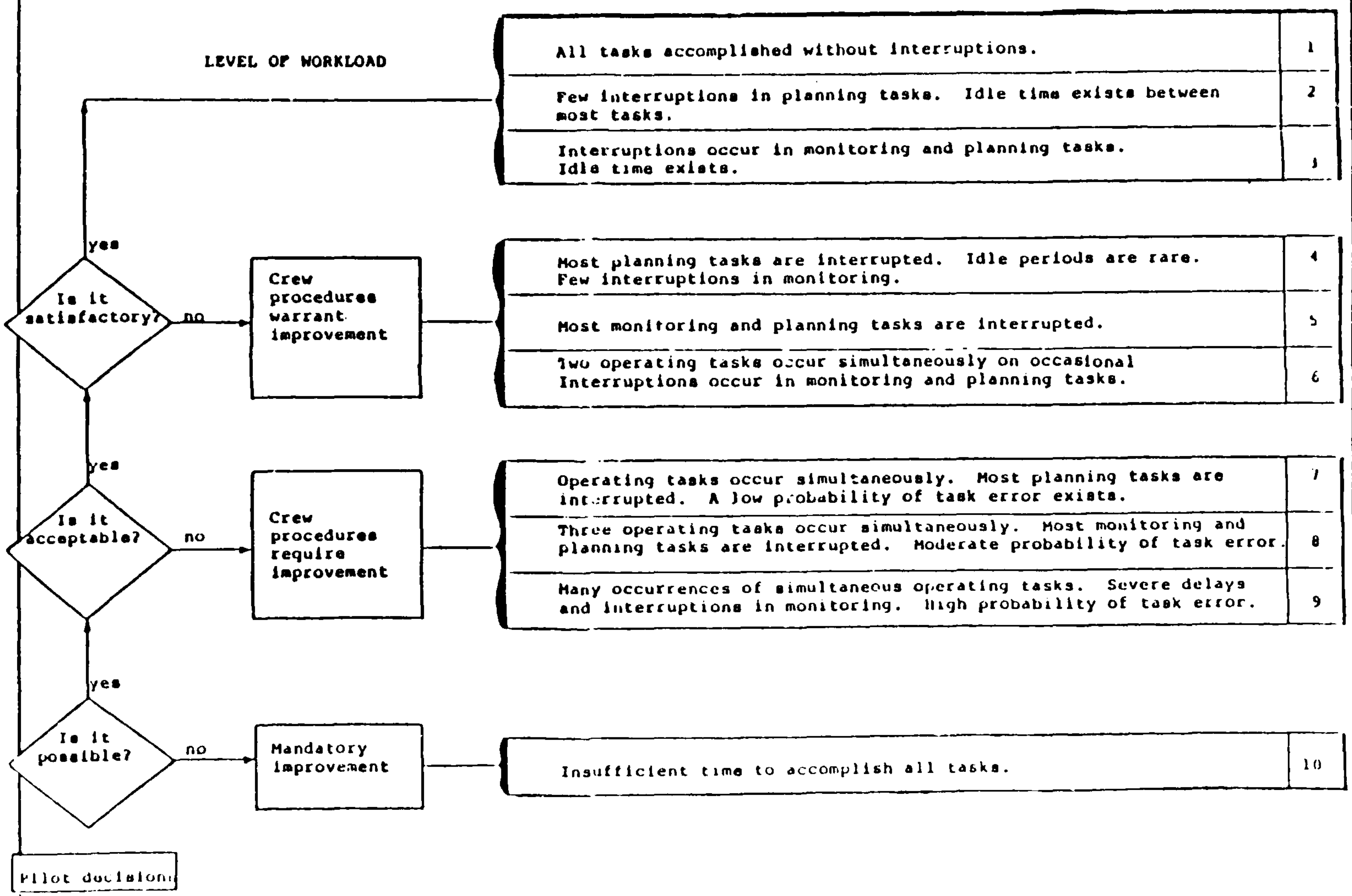
WORKLOAD INDEX	LATERAL ERROR	VELOCITY ERROR	ALTITUDE ERROR	GLOBAL HQ	GLOBAL DEMANDS	COPPER-HARPER
F-RATIO 22.46	16.95	1.40	1.50	18.42	18.57	39.28
CORRELATION WITH WORKLOAD						
1.00	0.907	0.692	0.474	0.812	0.804	0.879

$F_{0.01} = 3.48$, $F_{0.001} = 4.95$

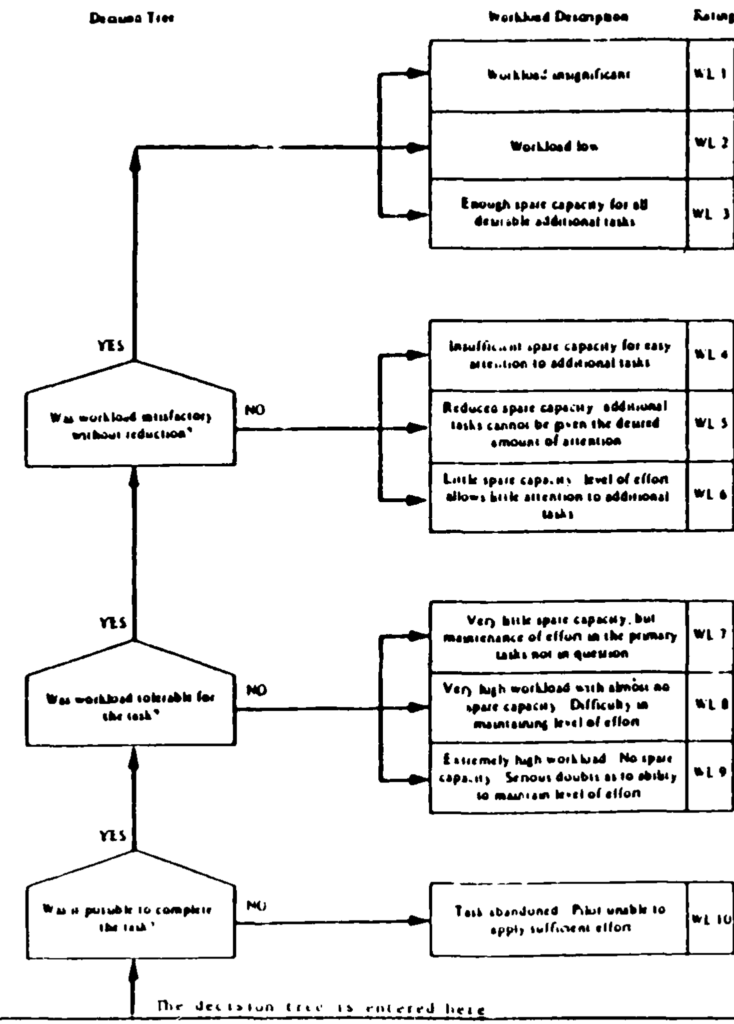
$R_{0.05} = 0.602$, $R_{0.01} = 0.735$

Stackhouse & Peterson, 1972

RATING SCALES: ORIGINAL SIMPSON-SHERIDAN SCALE



SUBJECTIVE RATINGS: BEDFORD SCALE



SUBJECTIVE RATINGS: BEDFORD SCALE

COMPARISON OF BEDFORD SCALE RATINGS WITH MEAN HEART RATE DURING CERTIFICATION FLIGHTS FOR THE BAE 146

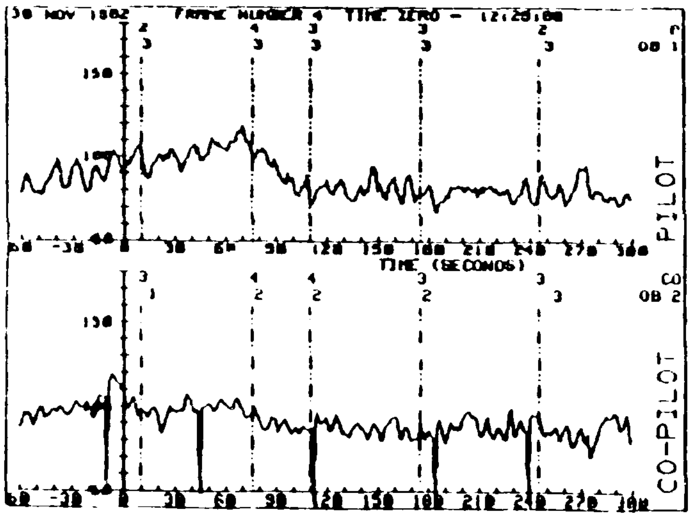
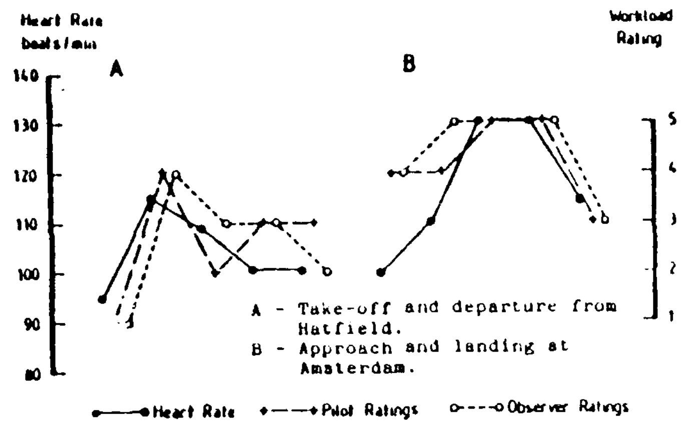
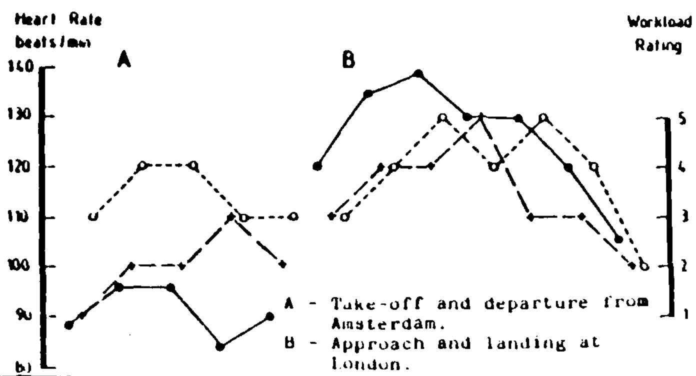


Fig 10 Example of beat-to-beat heart rate plots and workload ratings from the BAe 146 certification programme. (Recorded during a take-off from Amsterdam.)



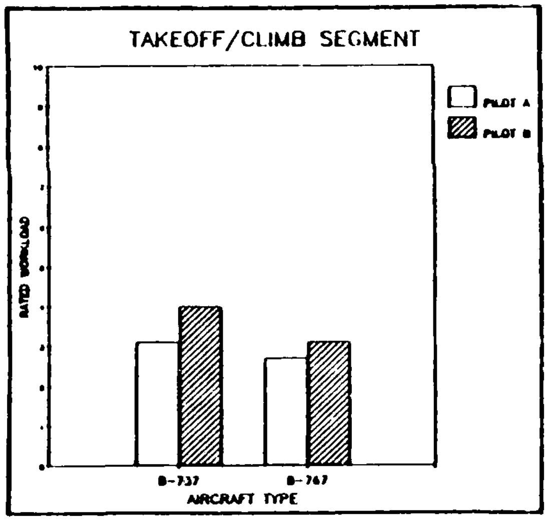


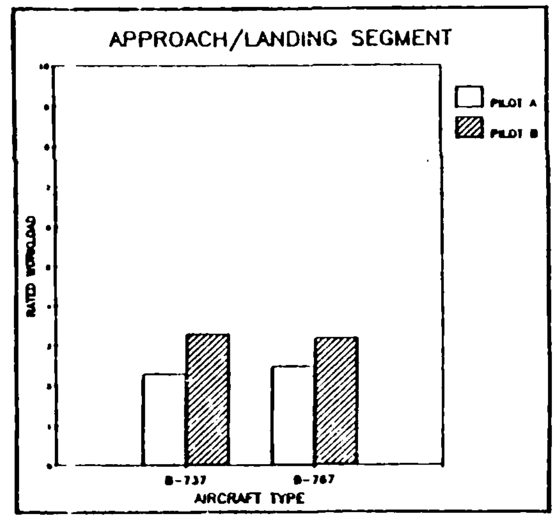
NOTE: OBSERVER RATINGS WERE MORE CLOSELY RELATED TO VARIATIONS IN HEART RATE THAN PILOTS'

Roscoe, 1984

RATING SCALES: MODIFIED COOPER-HARPER SCALE BEDFORD SCALE

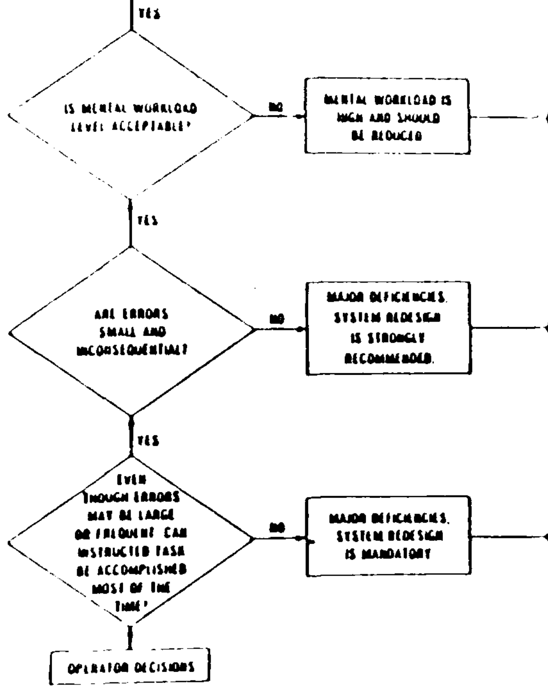
EXAMPLE: INFLIGHT RATINGS WERE OBTAINED FROM TWO PILOTS DURING MANY FLIGHTS TO COMPARE THE WORKLOAD EXPERIENCED IN THE B-737 AND THE B-767





Roscoe & Grieve

RATING SCALES: MODIFIED COOPER-HARPER SCALE (WIERWILLE)



DIFFICULTY LEVEL	OPERATOR DEMAND LEVEL	RATING
VERY EASY, HIGHLY DESIRABLE	OPERATOR MENTAL EFFORT IS MINIMAL AND DESIRED PERFORMANCE IS EASILY ATTAINABLE	1
EASY, DESIRABLE	OPERATOR MENTAL EFFORT IS LOW AND DESIRED PERFORMANCE IS ATTAINABLE	2
FAIR, MILD DIFFICULTY	ACCEPTABLE OPERATOR MENTAL EFFORT IS REQUIRED TO ATTAIN ADEQUATE SYSTEM PERFORMANCE	3
MINOR BUT ANNOYING DIFFICULTY	MODERATELY HIGH OPERATOR MENTAL EFFORT IS REQUIRED TO ATTAIN ADEQUATE SYSTEM PERFORMANCE	4
MODERATELY OBJECTIONABLE DIFFICULTY	HIGH OPERATOR MENTAL EFFORT IS REQUIRED TO ATTAIN ADEQUATE SYSTEM PERFORMANCE	5
VERY OBJECTIONABLE BUT TOLERABLE DIFFICULTY	MAXIMUM OPERATOR MENTAL EFFORT IS REQUIRED TO ATTAIN ADEQUATE SYSTEM PERFORMANCE	6
MAJOR DIFFICULTY	MAXIMUM OPERATOR MENTAL EFFORT IS REQUIRED TO BRING ERRORS TO MODERATE LEVEL	7
MAJOR DIFFICULTY	MAXIMUM OPERATOR MENTAL EFFORT IS REQUIRED TO AVOID LARGE OR NUMEROUS ERRORS	8
MAJOR DIFFICULTY	INTENSE OPERATOR MENTAL EFFORT IS REQUIRED TO ACCOMPLISH TASK, BUT FREQUENT OR NUMEROUS ERRORS PERSIST	9
IMPOSSIBLE	INSTRUCTED TASK CANNOT BE ACCOMPLISHED RELIABLY	10

SUBJECTIVE RATING: CONSTRUCT VALIDITY

EXAMPLE: LACK OF CONVERGING EVIDENCE AMONG MANY TYPES OF WORKLOAD MEASURES OBTAINED DURING FOUR EXPERIMENTS CONDUCTED IN A GAT SIMULATOR. INDIVIDUAL EXPERIMENTS EMPHASIZED PSYCHOMOTOR, PERCEPTUAL, MEDIATIONAL, AND COMMUNICATIONS SOURCES OF WORKLOAD. WHICH EVIDENCE DO YOU BELIEVE? WERE THE LEVELS OF DEMANDS IMPOSED BY TASKS WITHIN EACH EXPERIMENT REALLY DIFFFERENT?

	Experiment			
	Psychomotor	*Perceptual*	*Mediational*	*Communication*
Opinion Measures				
Cooper-Harper Scale	A	—	—	—
Modified Cooper-Harper Scale	—	B	B	B
Workload-Compensation-Interference/ Tech Effectiveness Scale	A	B	B	—
Multiple Descriptor Scale	—	B	ns	C
Primary Task Measures				
RMS Localizer Error	ns	—	—	—
RMS Glide Slope Error	ns	—	—	—
Control Movements per Second	A	ns	ns	ns
Pitch High Pass Mean Square	—	ns	ns	ns
Roll High Pass Mean Square	—	ns	ns	ns
Perceptual Response Time	—	A	—	—
Mediational Response Time	—	—	A	—
Communications Response Time	—	—	—	C(REV)
Mediational Error Rate	—	—	B	—
Communications Errors of Omission	—	—	—	B
Communications Errors of Commission	—	—	—	B

	Experiment			
	Psychomotor	*Perceptual*	*Mediational*	*Communication*
Secondary Task Measures				
Digit Shadowing	ns	—	—	—
Memory Scanning (Sternberg)	ns	—	—	—
Mental Arithmetic	ns	—	—	—
Time Estimation Mean	ns	—	—	—
Standard Dev	B(NM)	C	B(INT)	B
Absolute Error	ns	—	—	—
RMS Error	ns	—	—	—
Tapping Regularity	—	B	ns	ns
Physiological Measures				
Pulse Rate Mean	C	ns	ns	ns
Pulse Rate Standard Deviation	ns		ns	ns
Pupil Diameter	ns	ns	ns	B(NM)
Respiration Rate	ns	B		
Eye Blinks per Second	ns	ns	B(NM)	ns
Eye Transition Frequency	ns	—	—	—
Eye Fixations	—	—	B	ns
Voice Pattern Analysis	ns	—	—	—

KEY

A, B, C (Discriminated among 3, 2, or 1 pairs of load conditions

ns (Did not discriminate among any conditions)

— (The measure was not included

INT, NM, (The measure was interferring or non-monotonically related with load levels

Wierwille, Casali, Connor, & Rahimi, 1985

MULTI-DIMENSIONAL SCALES

SUBJECTIVE RATINGS: WORKLOAD/COMPENSATION/INTERFERENCE/ TECHNICAL EFFECTIVENESS SCALE

TECHNICAL EFFECTIVENESS	1	2	3	4 Column	Row
Multiple Tasks Integrated					4
Design Enhances Specific Task Accomplishment					3
Adequate Performance Achievable; Design Sufficient to Specific Task					2
Inadequate performance Due to Technical Design					1
	Workload Extreme; Compensation Extreme; Interference Extreme	Workload High; Compensation High; Interference High	Workload Moderate; Compensation Moderate; Interference Moderate	Workload Low; Compensation Low; Interference Low	

RATING SCALE: LATER VERSION OF SIMPSON-SHERIDAN SCALE

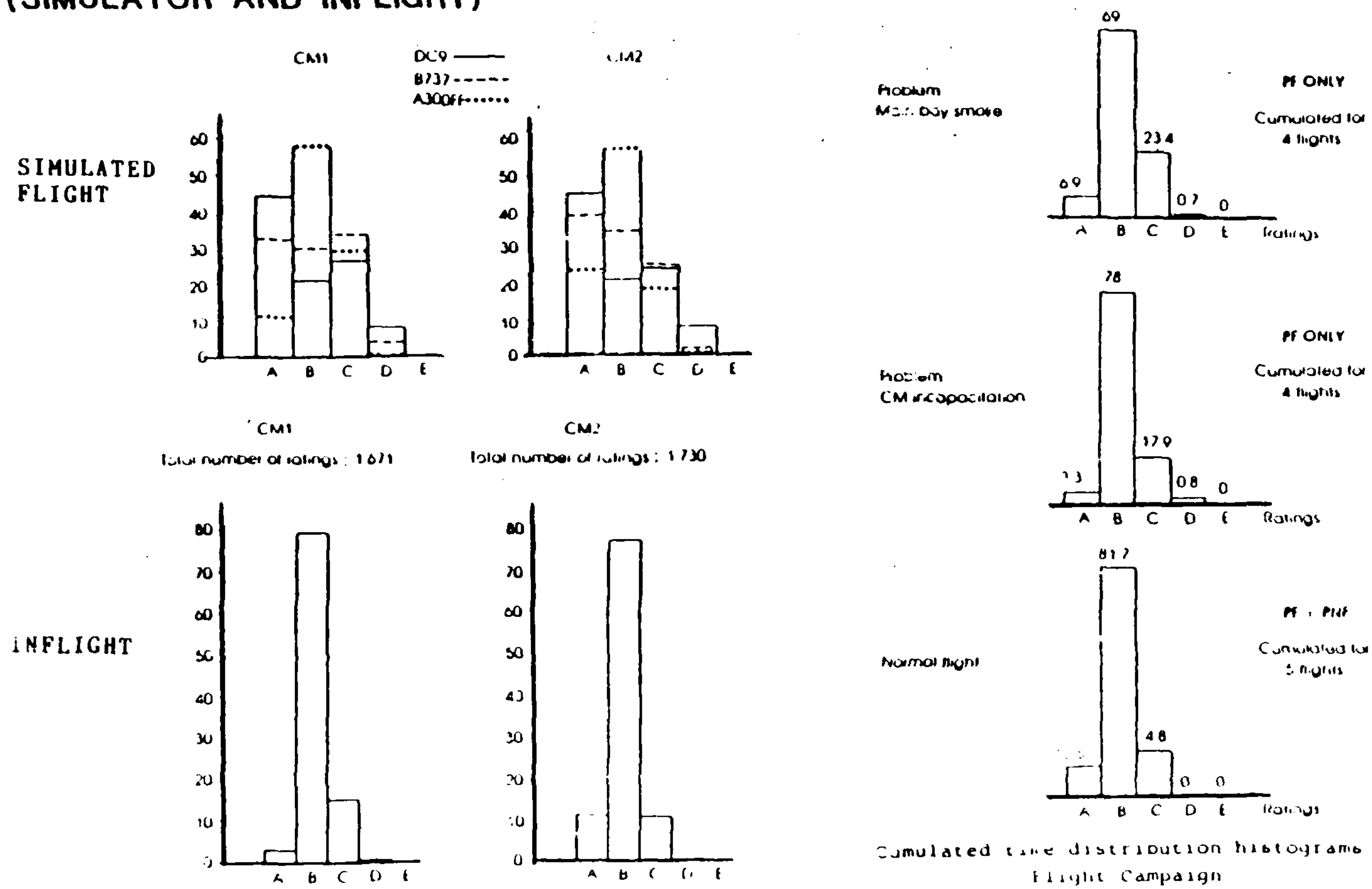
1	FRACTION OF TIME BUSY		
	seldom have anything to do		
	have free moments of time:	often	
		occasionally	
		very rarely	
	fully occupied every single instant		
2.	INTENSITY OF THINKING/INFORMATION-PROCESSING		
	[illegible] is completely automatic; no conscious thinking or planning required		
	[illegible] effort and planning required due to problem complexity, uncertainty, unpredictability, [illegible], etc., is:	low level, occasional	
		moderate	
		high level	
	supreme mental effort and concentration are absolutely necessary		
3.	INTENSITY OF FEELING		
	experience is relaxing, nothing to be concerned about		
	emotional stress, anxiety, worry, frustration, confusion, etc., are:	mild, occasional	
		moderate	
		high level	
	severe and intense psychological stress		

SUBJECTIVE RATINGS: AIRBUS RATING SCALE

WORKLOAD ASSESSMENT		CRITERIA			APPRECIATION
		RESERVE CAPACITY	INTERRUPTIONS	EFFORT OR STRESS	
LIGHT	2	AMPLE	—	—	VERY ACCEPTABLE
MODERATE	3	ADEQUATE	SOME	—	WELL ACCEPTABLE
FAIR	4	SUFFICIENT	RECURRING	NOT UNDUE	ACCEPTABLE
HIGH	5	REDUCED	REPETITIVE	MARKED	HIGH BUT ACCEPTABLE
HEAVY	6	LITTLE	FREQUENT	SIGNIFICANT	JUST ACCEPTABLE
EXTREME	7	NONE	CONTINUOUS	ACUTE	NOT ACCEPTABLE CONTINUOUSLY
SUPPREME	8	IMPAIRMENT	IMPAIRMENT	IMPAIRMENT	NOT ACCEPTABLE INSTANTEANOUSLY

SUBJECTIVE RATINGS: AIRBUS RATING SCALE

SUMMARY OF WORKLOAD RATINGS OBTAINED DURING A300FF CERTIFICATION (SIMULATOR AND INFLIGHT)

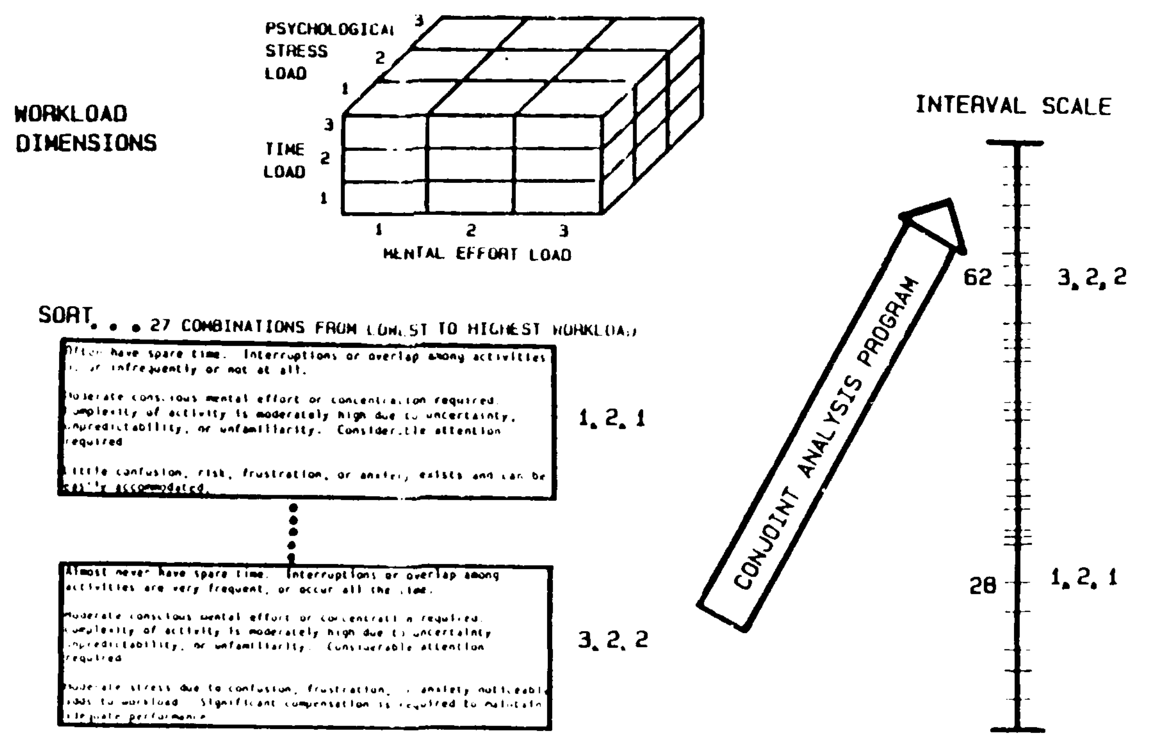


Speyer & Fort, 1982

SUBJECTIVE WORKLOAD ASSESSMENT TECHNIQUE

SWAT

SUBJECTIVE RATINGS: SUBJECTIVE WORKLOAD ASSESSMENT TECHNIQUE (SWAT)



SUBJECTIVE RATINGS: SWAT (EVIDENCE OF SENSITIVITY)

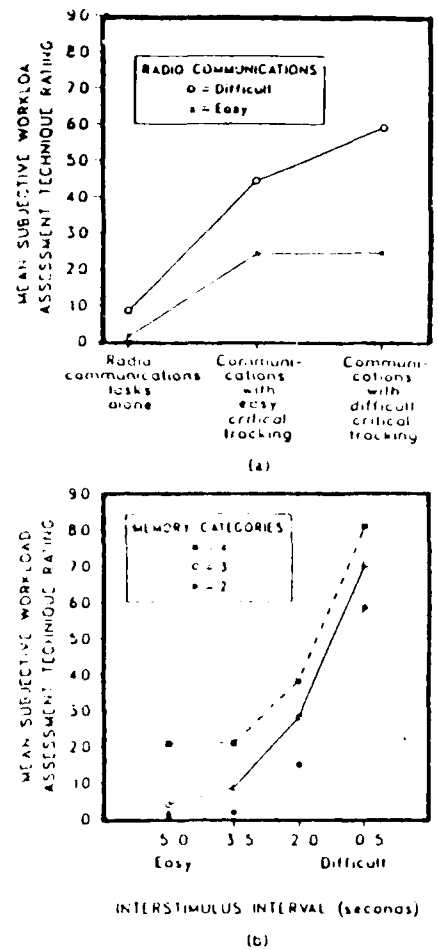


Figure 42.8 Mean Subjective Workload Assessment Technique (SWAT) ratings as a function of task difficulty in several different types of tasks. (a) The effects of two levels of primary task tracking difficulty with a simple and difficult version of a secondary aircrew radio communications task (Reid, Shingledecker, & Eggemeier, 1981). Both radio communications condition and tracking task difficulty significantly affected ($p < .01$) mean SWAT ratings. Post hoc multiple comparisons tests indicated that low difficulty tracking ratings were significantly different from those associated with high difficulty tracking ($p < .01$), and that ratings from the single task conditions were lower than ratings from the dual task conditions. SWAT ratings, therefore, distinguished levels of difficulty in the tracking task and reflected the additional demands imposed by the more difficult dual task condition. (b) The effects of variations in memory task difficulty on mean SWAT ratings (Eggemeier, Crabtree, Zingg, Reid, & Shingledecker, 1982). Difficulty of the memory task was manipulated by varying the number of information categories (two, three, or four) that were to be retained in memory and the interstimulus interval (0.5, 2.0, 3.5, and 5.0 seconds). Both manipulations produced significant effects ($p < .01$) on SWAT ratings, thereby supporting the sensitivity of the procedure to difficulty manipulations in this task. Taken together, the results of Graphs (a) and (b) support the capability of the SWAT technique to discriminate workload differences in both central processing and motor output tasks. (Redrawn from G. B. Reid, C. A. Shingledecker, & F. T. Eggemeier, Application of conjoint measurement to workload scale development. *Proceedings of the Human Factors Society 25th Annual Meeting*. Copyright 1981 by Human Factors Society. And redrawn from F. T. Eggemeier, G. B. Reid, & C. A. Shingledecker, Subjective workload assessment in a memory update task. *Proceedings of the Human Factors Society 26th Annual Meeting*. Copyright 1982 by Human Factors Society. Reprinted with permission.)

SUBJECTIVE RATINGS: SWAT (CONVERGING EVIDENCE OF VALIDITY)

EXAMPLE: SWAT, NASA BIPOLAR, AND HQR RATINGS WERE OBTAINED DURING AN ADVANCED HELICOPTER SIMULATION. LEVEL OF AUTOMATION AND CREW SIZE (ONE vs TWO) WERE VARIED.

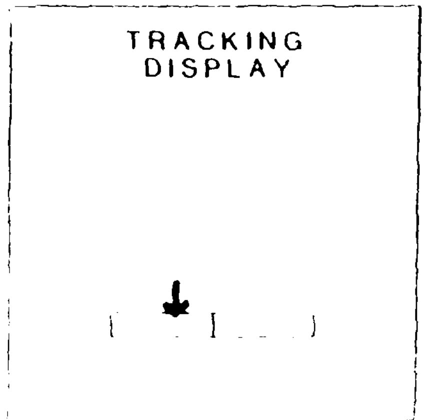
SWAT vs NASA BIPOLAR RATINGS

SWAT vs COOPER-HARPER HQRs

NOTE: THESE DATA REPRESENT THE NOE SEGMENT ONLY

Haworth, Bivens, & Shively, 1986

SUBJECTIVE RATINGS: SWAT vs NASA BIPOLAR RATINGS
COMPARISON OF RELIABILITY/SENSITIVITY (1)

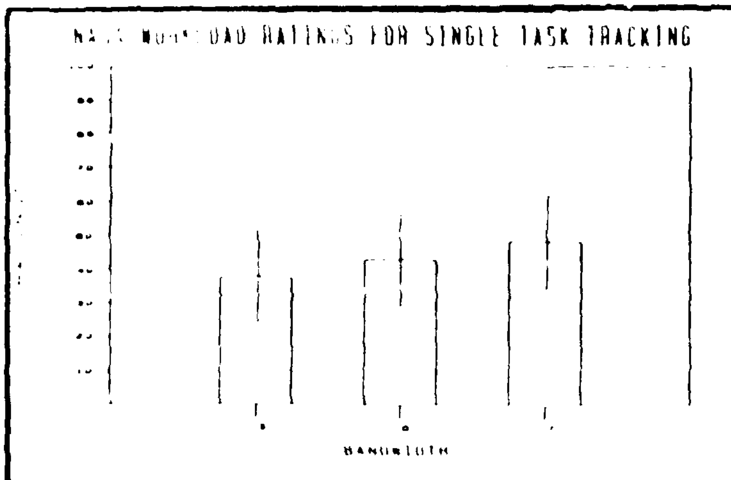


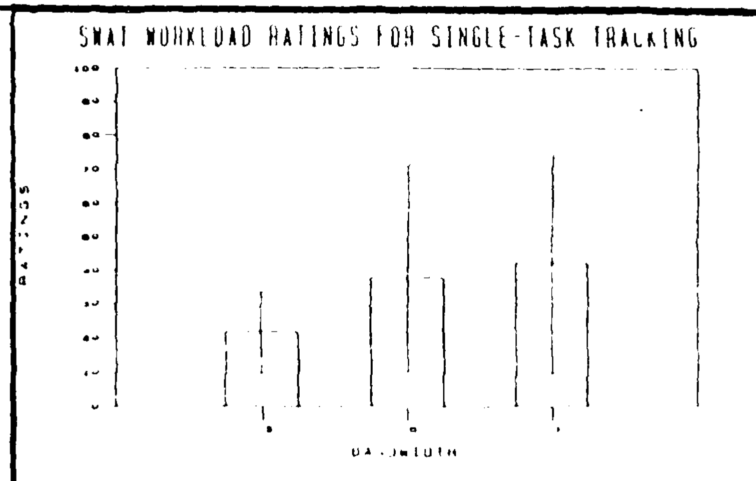
OBJECTIVE: TO COMPARE THE SENSITIVITY AND STABILITY OF TWO MULTI-DIMENSIONAL RATING TECHNIQUES FOR WORKLOAD ASSESSMENT

APPROACH: TRACKING TASKS
RATINGS WERE OBTAINED WITH EACH TECHNIQUE FOLLOWING THE PERFORMANCE OF A SINGLE-AXIS TRACKING TASK:
BANDWIDTHS = .3, .5, .7

RESULTS: ALTHOUGH BOTH MEASURES INDICATED INCREASING WORKLOAD AS BANDWIDTH INCREASED, THE DIFFERENCE WAS STATISTICALLY SIGNIFICANT FOR THE NASA RATINGS ONLY.

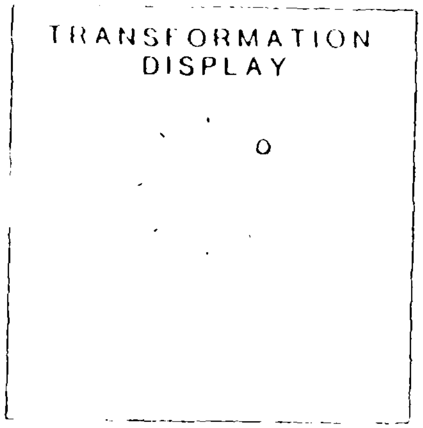
BETWEEN-SUBJECT VARIABILITY WAS REDUCED MORE BY THE NASA TECHNIQUE THAN BY SWAT

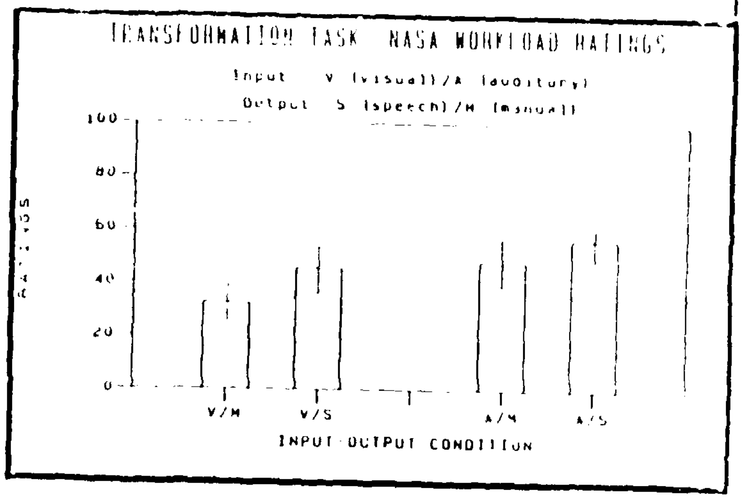


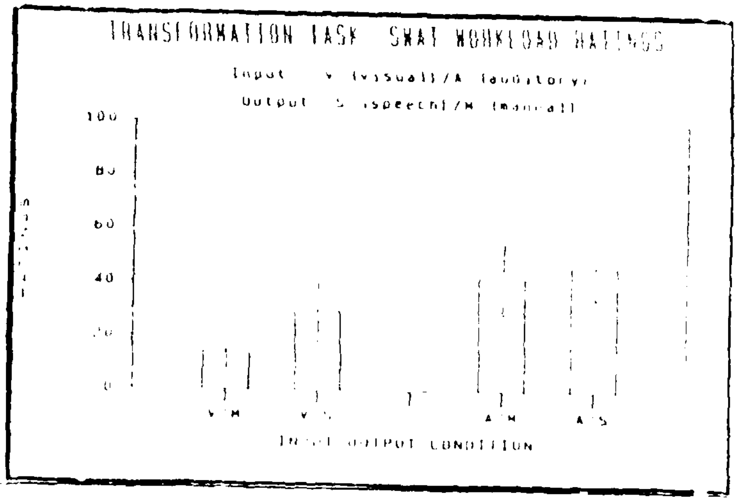


Vidulich & Tsang, 1985

SUBJECTIVE RATINGS: SWAT vs NASA BIPOLAR RATINGS
COMPARISON OF RELIABILITY/SENSITIVITY (2)







OBJECTIVE: TO COMPARE THE SENSITIVITY AND STABILITY OF TWO, MULTI-DIMENSIONAL RATING TECHNIQUES FOR WORKLOAD ASSESSMENT

APPROACH: TRANSFORMATION TASKS
RATINGS WERE OBTAINED WITH EACH TECHNIQUE FOLLOWING THE PERFORMANCE OF A SPATIAL TRANSFORMATION TASK:
(E.G. "NORTH-EAST", "SOUTH", etc)

RESULTS:

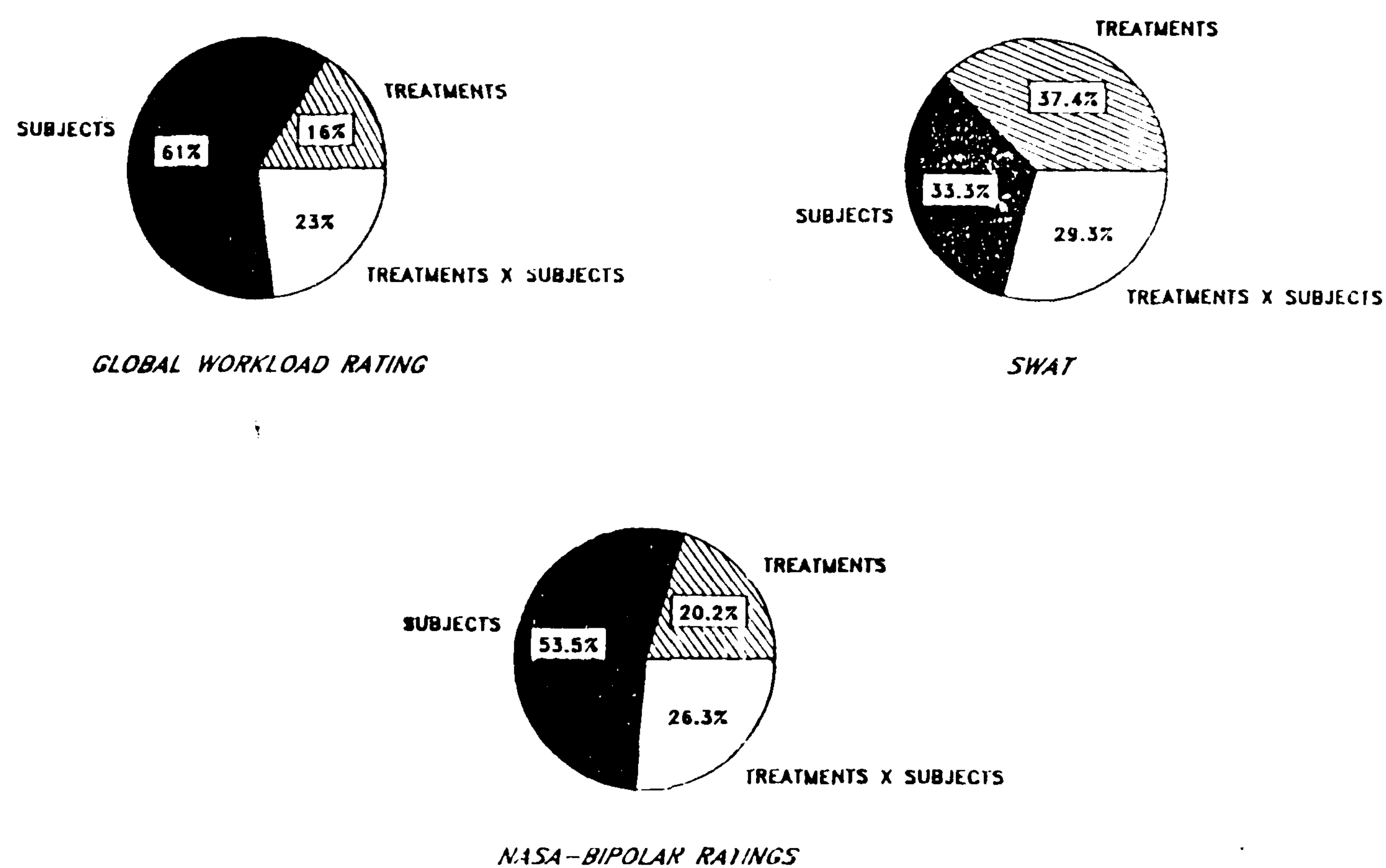
FOR BOTH TECHNIQUES, AUDITORY INPUT WAS RATED AS SIGNIFICANTLY MORE LOADING THAN VISUAL

FOR BOTH TECHNIQUES, SPEECH OUTPUT WAS RATED AS SIGNIFICANTLY MORE LOADING THAN MANUAL

AGAIN, BETWEEN-SUBJECT VARIABILITY WAS GREATER WITH THE SWAT TECHNIQUE

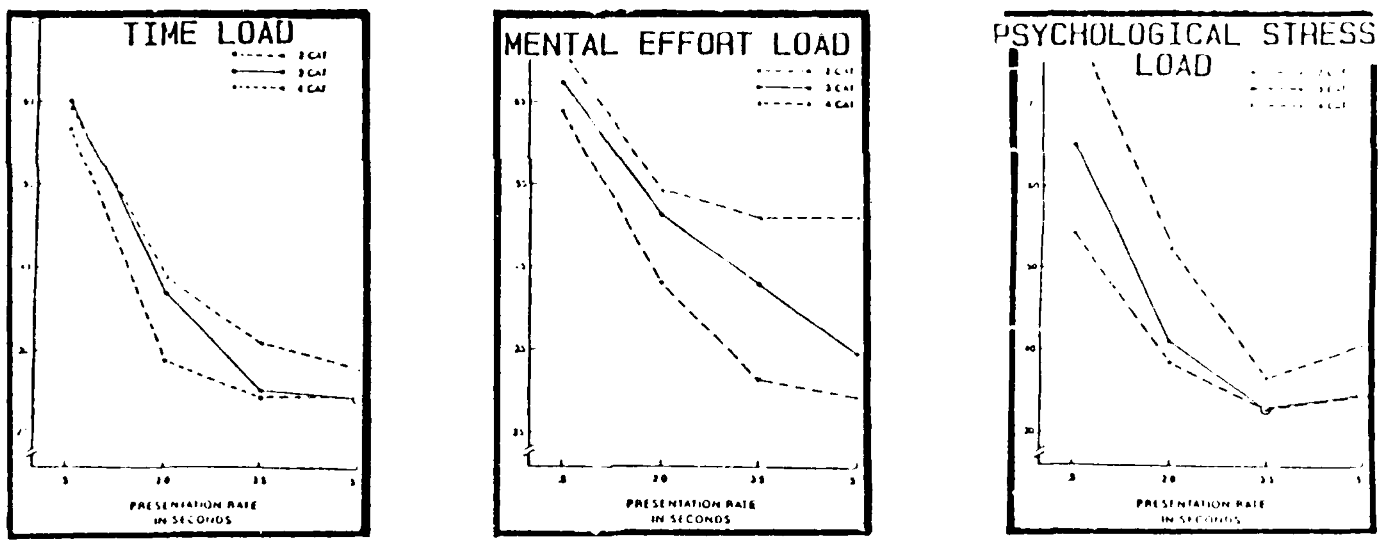
Vidulich & Tsang, 1986

COMPARISONS AMONG WORKLOAD RATINGS SOURCES OF VARIABILITY



Vidulich & Tsang 1986

SUBJECTIVE RATINGS: SWAT (EVIDENCE OF DIAGNOSTICITY)



NASA BIPOLAR RATING SCALE

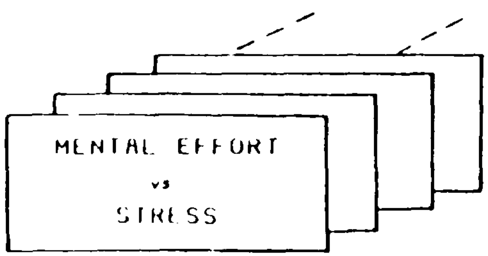
SUBJECTIVE RATINGS: NASA BIPOLAR RATING SCALE (EARLY VERSION OF NASA-TLX)

WORKLOAD DIMENSIONS:

- TASK DIFFICULTY
- TIME PRESSURE
- OWN PERFORMANCE
- PHYSICAL EFFORT
- MENTAL EFFORT
- STRESS
- FATIGUE
- FRUSTRATION
- TYPE OF TASK

"WEIGHTS"

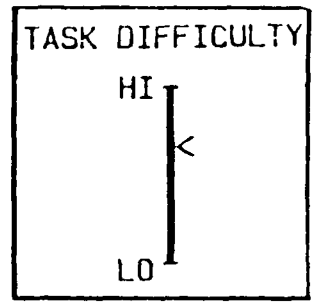
EACH OF 9 FACTORS IS COMPARED WITH EVERY OTHER ONE (WHICH IS MORE RELATED TO WORKLOAD?)



0 = NEVER SELECTED
8 = ALWAYS SELECTED

BIPOLAR RATINGS:

THE AMOUNT OF EACH FACTOR EXPERIENCED IN A TASK IS EVALUATED ON A BIPOLAR SCALE:



WEIGHTING PROCEDURE

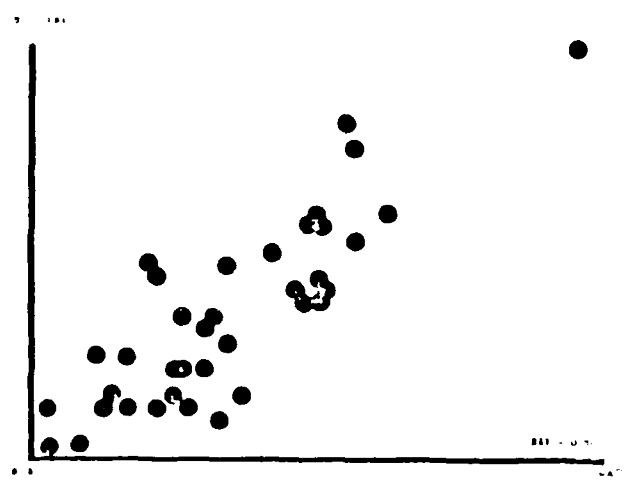
EACH "RATING" IS WEIGHTED BY ITS SUBJECTIVE IMPORTANCE TO EACH SUBJECT (THE "WEIGHTS")

THE AVERAGE OF THE WEIGHTED RATINGS = DERIVED WORKLOAD SCORE

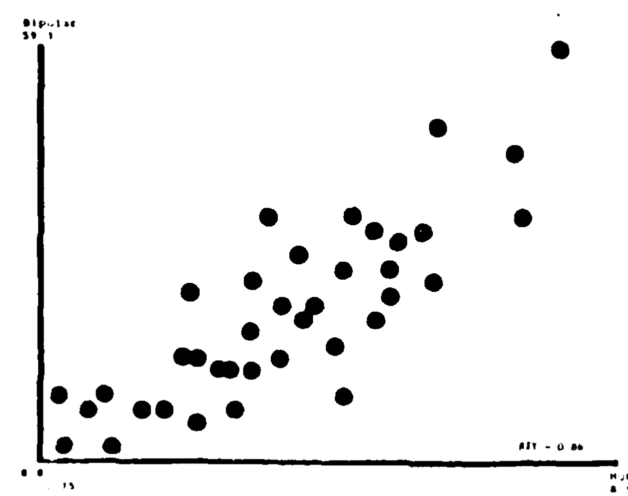
SUBJECTIVE RATINGS: NASA BIPOLAR RATINGS (CONVERGING EVIDENCE OF VALIDITY)

EXAMPLE: SWAT, NASA BIPOLAR, AND HQR RATINGS WERE OBTAINED DURING AN ADVANCED HELICOPTER SIMULATION. LEVEL OF AUTOMATION AND CREW SIZE (ONE vs TWO) WERE VARIED.

NASA BIPOLAR vs SWAT



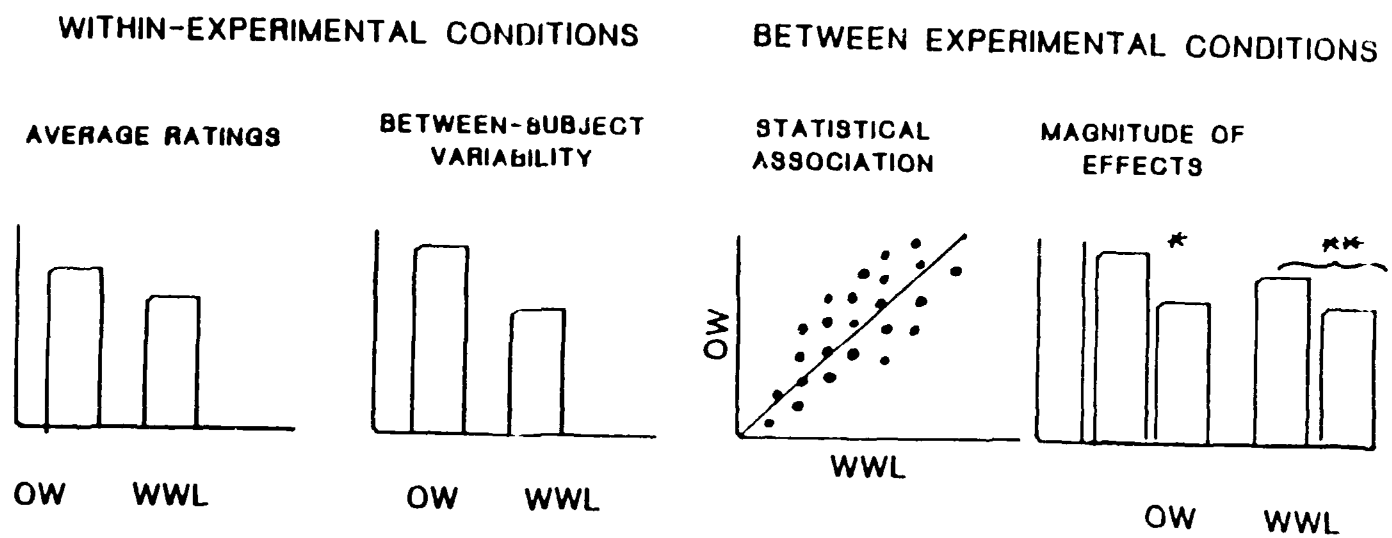
NASA BIPOLAR vs COOPER-HARPER HQRs



NOTE: THESE DATA REPRESENT THE NOE SEGMENT ONLY

Haworth, Bivens, & Shively, 1986

COMPARISON BETWEEN OVERALL WORKLOAD RATINGS AND WEIGHTED WORKLOAD SCORE



SUMMARY OF NASA BIPOLAR RATING SCALE EVALUATION

- SOURCES OF WORKLOAD VARY BETWEEN DIFFERENT TYPES OF TASKS
- RATINGS OF COMPONENT FACTORS ARE MORE DIAGNOSTIC THAN GLOBAL WORKLOAD RATINGS
- SUBJECTIVE WORKLOAD DEFINITIONS DO VARY (THEREBY CONTRIBUTING TO BETWEEN-SUBJECT VARIABILITY)
- HOWEVER, THEIR A PRIORI BIASES ABOUT WORKLOAD ARE UNRELATED TO THEIR RATINGS OF WORKLOAD AND WORKLOAD COMPONENTS
- SUBSCALES:
  - o SOME WERE HIGHLY CORRELATED (E.G. STRESS, FRUSTRATION)
    - THEY WERE COMBINED
  - o OTHERS WERE UNRELATED TO WORKLOAD (E.G. FATIGUE)
    - THEY WERE DELETED
  - o OTHERS WERE TOO BROAD (E.G., TASK DIFFICULTY)
    - THEY WERE SUBDIVIDED

NASA-TASK LOAD INDEX

NASA-TLX

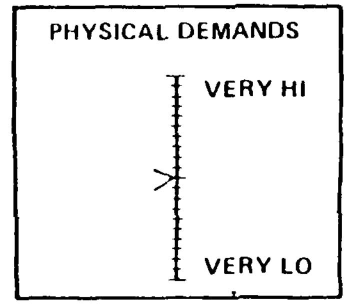
RATING SCALES: NASA-TASK LOAD INDEX (NASA-TLX)

WORKLOAD FACTORS

- PHYSICAL DEMANDS (PD)
- MENTAL DEMANDS (MD)
- TEMPORAL DEMANDS (TD)
- OWN PERFORMANCE (OP)
- EFFORT (EF)
- FRUSTRATION (FR)

BIPOLAR RATINGS

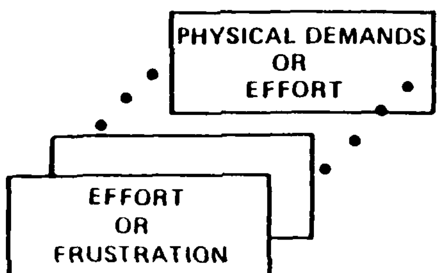
THE MAGNITUDE OF EACH FACTOR IS EVALUATED ON A BIPOLAR SCALE



FACTOR WEIGHTS

EACH FACTOR COMPARED TO EVERY OTHER:

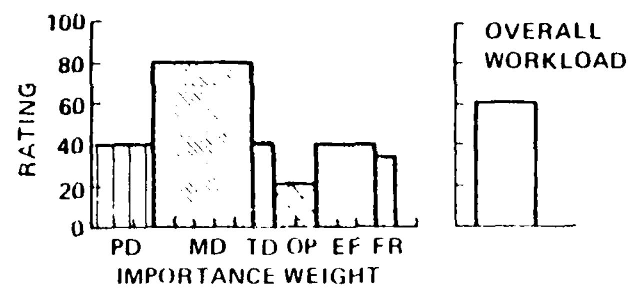
WHICH MEMBER OF EACH PAIR CONTRIBUTED MORE TO THE WORKLOAD OF THE TASK?



0 = NEVER SELECTED
5 = ALWAYS SELECTED

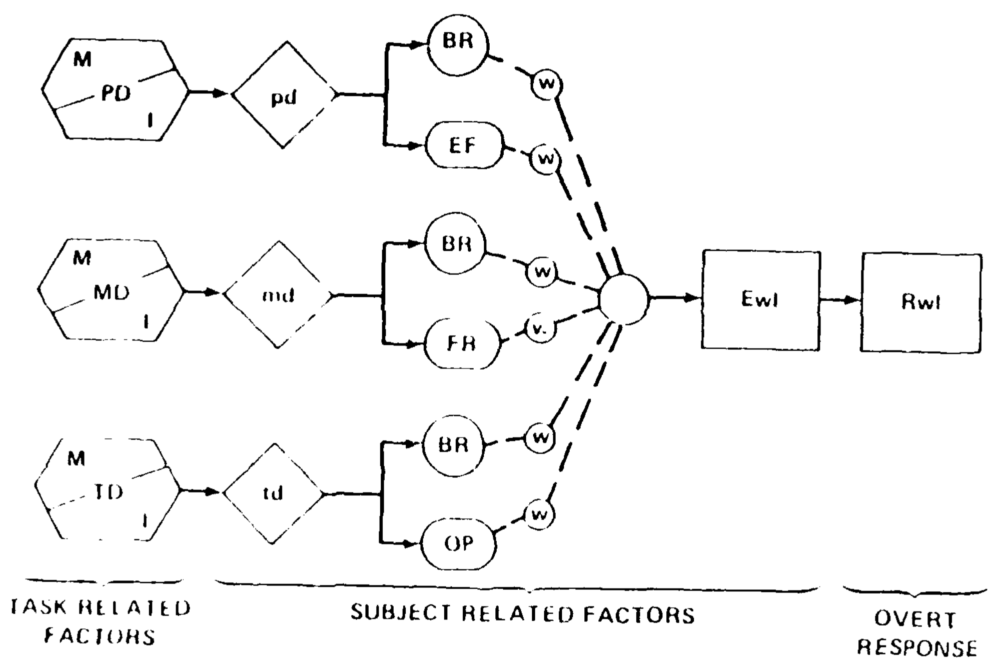
WEIGHTED RATINGS

OVERALL WORKLOAD (OW) = MEAN OF WEIGHTED RATINGS



RATING SCALES: NASA-TASK LOAD INDEX (NASA-TLX)

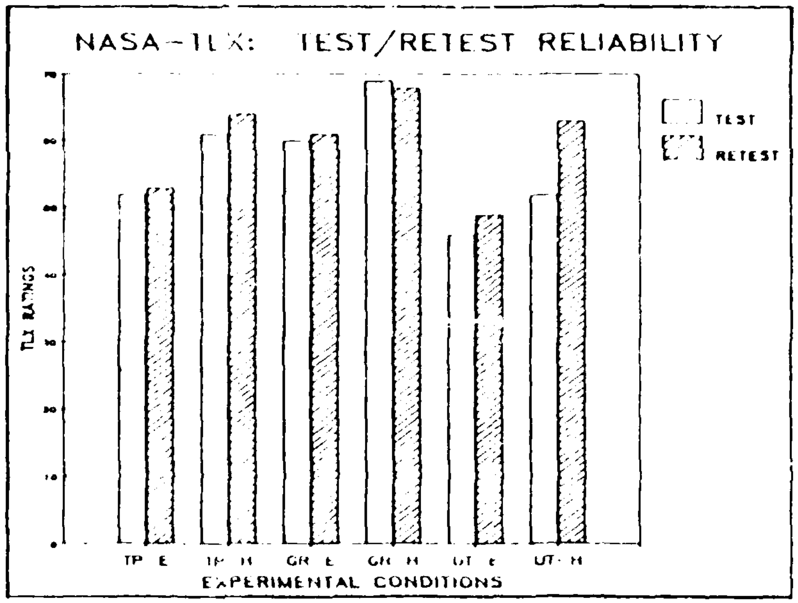
MODEL OF SUBJECTIVE WORKLOAD ESTIMATION PROCESS



PD, MD, TD	Objective physical, mental and temporal task demands
M, I	Objective magnitudes and importance of sources of demands
pd, md, td	Psychological representations of task demands
BR	Behavioral responses to task demands
OP, EF, FR	Subjective responses/evaluations of behavioral responses
w	Subjective weighting of factors
Ewl	Integrated subjective experience of workload
Rwl	Formal numeric or verbal evaluation of workload

SUBJECTIVE RATINGS: NASA-TASK LOAD INDEX (NASA-TLX) TEST/RETEST RELIABILITY

SUBJECTS PROVIDED TLX RATINGS FOLLOWING ASYMPTOTIC PERFORMANCE OF TWO LEVELS (E,H) OF THREE TASKS (TIMEPOOLS TARGET ACQUISITION, GRAMMATICAL REASONING, AND UNSTABLE TRACKING) TWICE, SEPARATED BY AN INTERVAL OF 4 WEEKS. NO SIGNIFICANT DIFFERENCES WERE FOUND.

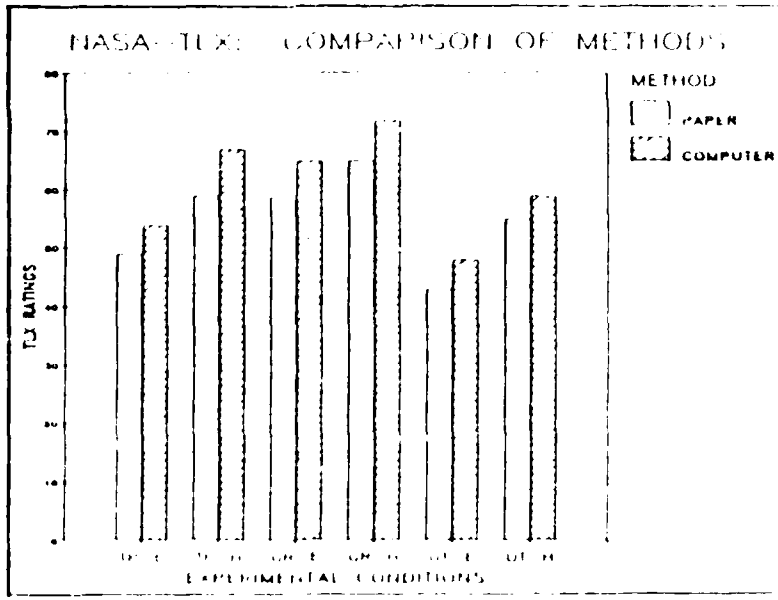


NOTE: CORRELATION IS .83

King, 1987

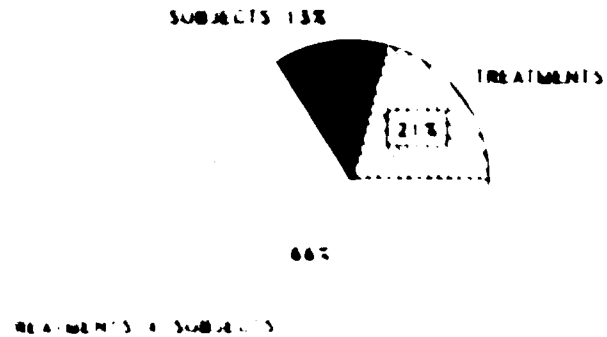
SUBJECTIVE RATINGS: NASA-TASK LOAD INDEX (NASA-TLX) COMPARISON OF ALTERNATIVE FORMS

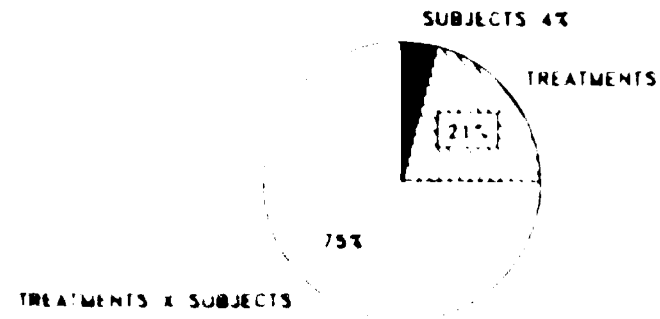
SUBJECTS PROVIDED TLX RATINGS FOLLOWING ASYMPTOTIC PERFORMANCE OF TWO LEVELS (E,H) OF THREE TASKS (TIMEPOOLS TARGET ACQUISITION, GRAMMATICAL REASONING, AND UNSTABLE TRACKING) USING TWO FORMS OF THE SCALE: PAPER AND PENCIL AND COMPUTERIZED. NO SIGNIFICANT DIFFERENCES WERE FOUND, ALTHOUGH THE COMPUTER RATINGS WERE 5.7 POINTS HIGHER, ON THE AVERAGE.



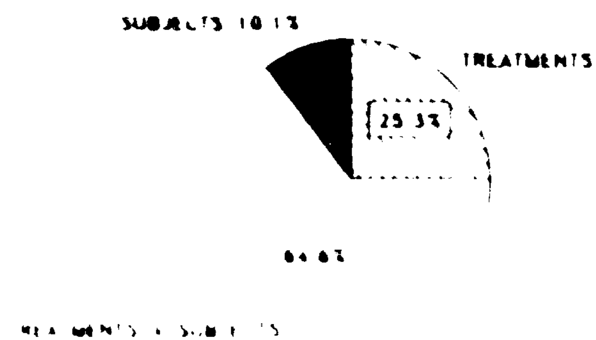
NOTE: CORRELATION = .94

NASA TLX: COMPARISON BETWEEN METHODS OF PRESENTATION (VARIABILITY DISTRIBUTIONS)





PAPER AND PENCIL VERSION



SUBJECTIVE RATINGS: NASA-TASK LOAD INDEX (NASA-TLX)
EMPIRICAL CORRELATIONS AMONG SUBSCALE RATINGS

[illegible] LABORATORY TASKS. ALTHOUGH
[illegible] WITH A SINGLE RATING OF
[illegible] INDEPENDENT INFORMATION

[illegible] among bipolar ratings

	MD	PD	TD	OP	EF	FR
[illegible]	[illegible]					
[illegible]	[illegible]	50				
[illegible]	[illegible]	27	32			
[illegible]	[illegible]	58	66	40		
[illegible]	[illegible]	11	[illegible]	57	69	
[illegible]	[illegible]	[illegible]	67	[illegible]	81	70

Hart & Staveland, in press

SUBJECTIVE RATINGS: NASA-TASK LOAD INDEX (NASA-TLX) SUBSCALE RATINGS AS PREDICTORS OF OVERALL WORKLOAD

RATINGS WERE OBTAINED FOR MANY LEVELS OF 20 LABORATORY TASKS. DEPENDING ON THE NATURE OF A TASK, DIFFERENT PATTERNS OF BETA WEIGHTS PROVIDED THE BEST PREDICTION OF OVERALL WORKLOAD. THIS SUGGESTS THAT WORKLOAD SOURCES AS WELL AS MAGNITUDES VARY FROM ONE ACTIVITY TO THE NEXT.

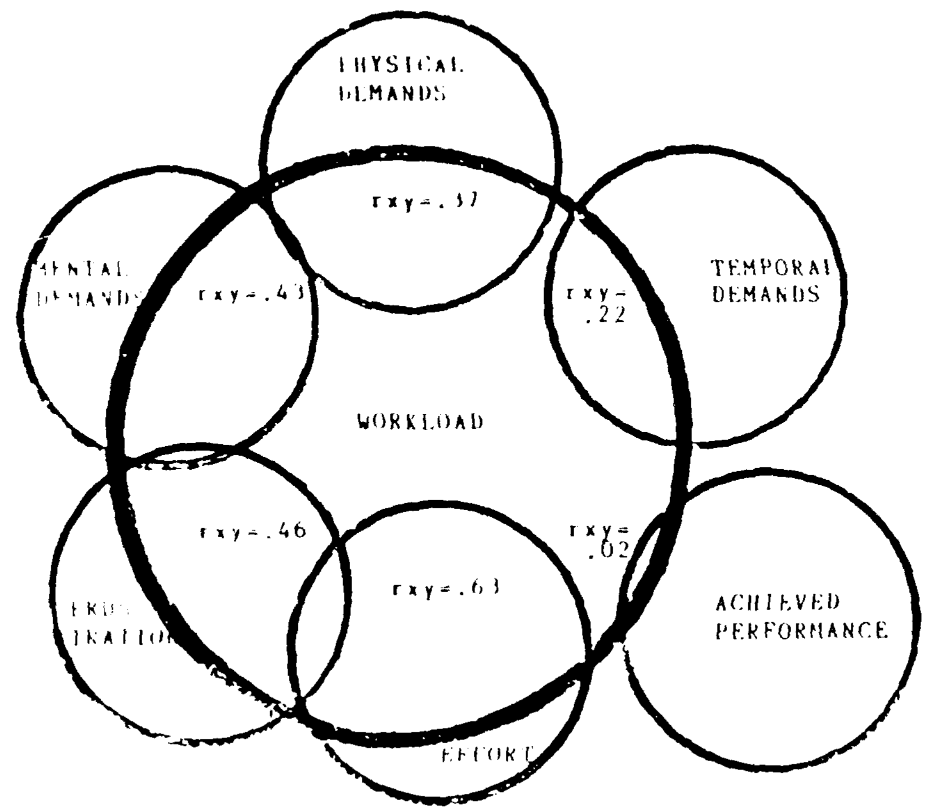
	r2	MD	PD	TD	OP	EF	FR
SINGLE-COGNITIVE	.88	.43*	.15*	.04	.01	.33*	.13*
SINGLE-MANUAL	.78*	.38*	.39*	.11*	.12*	.21*	.00
DUAL-TASKS	.82	.41*	.19*	.02	.09*	.29*	.20*
FITTSBERG	.86	.32*	.24*	.17*	.09*	.16*	.19*
POPCORN	.90	.34*	.23*	.22*	.03	.19*	.10*
OVERALL	.86	.38*	.22*	.08	.05	.24*	.16*

* $p < .01$

Hart & Staveland, in press

SUBJECTIVE RATINGS: NASA-TASK LOAD INDEX (NASA-TLX)
CORRELATIONS BETWEEN SUBSCALE RATINGS AND GLOBAL WORKLOAD RATINGS

ONE REASON WHY NASA TLX RATINGS PROVIDE A BETTER ESTIMATE OF WORKLOAD THAN A SINGLE GLOBAL MEASURE IS THAT THEY ARE THE AVERAGE OF SIX EVALUATIONS OF FACTORS THAT ARE EACH CORRELATED WITH WORKLOAD.

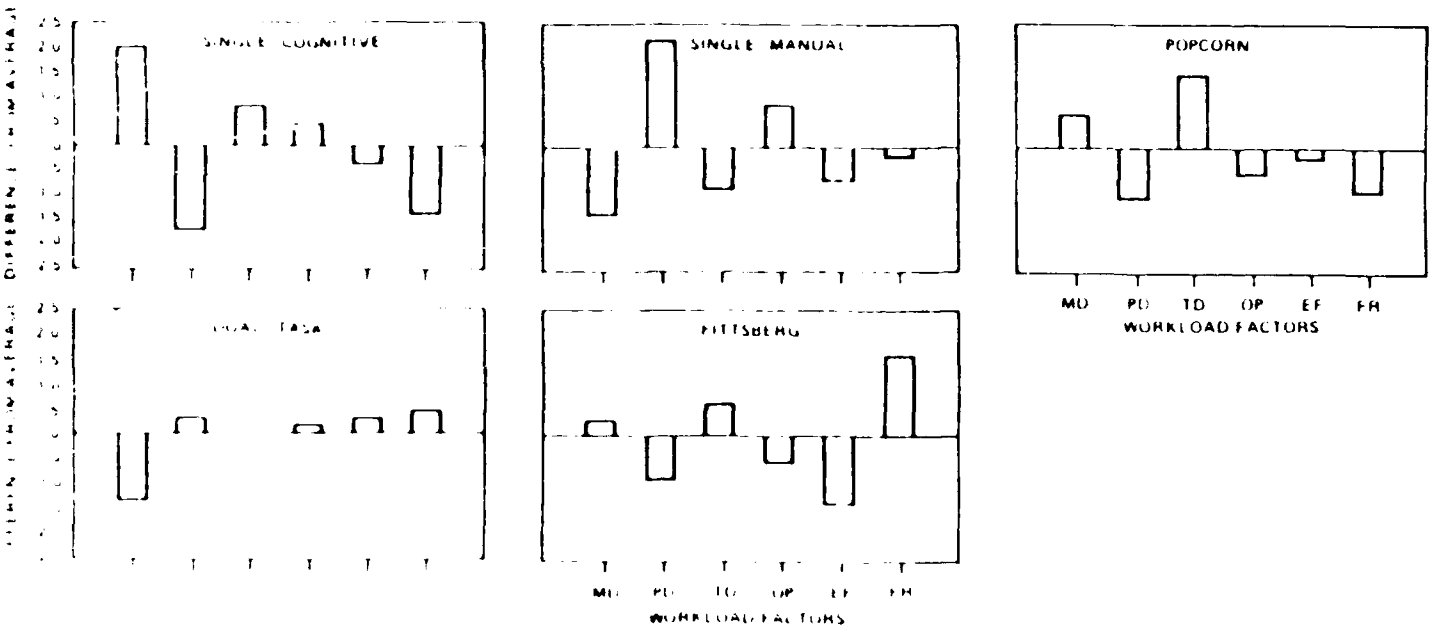


NOTE: CORRELATIONS AMONG SOME OF THE SUBSCALES WERE FOUND, AS WELL.

King, 1987

SUBJECTIVE RATINGS: NASA-TASK LOAD INDEX (NASA-TLX) SUBSCALE RATINGS AS PREDICTORS OF OVERALL WORKLOAD

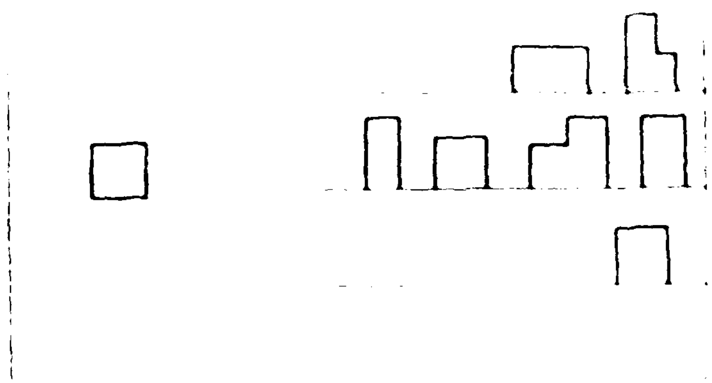
IMPORTANCE WEIGHTS WERE OBTAINED FOR EACH OF THE 20 TASKS. DEPENDING ON THE NATURE OF A TASK, DIFFERENT PATTERNS OF IMPORTANCE WEIGHTS WERE OBTAINED. THEIR DIFFERENT PATTERNS SUGGEST (AS DO THE RATINGS THEMSELVES) THAT WORKLOAD SOURCES VARY FROM ONE ACTIVITY TO THE NEXT.

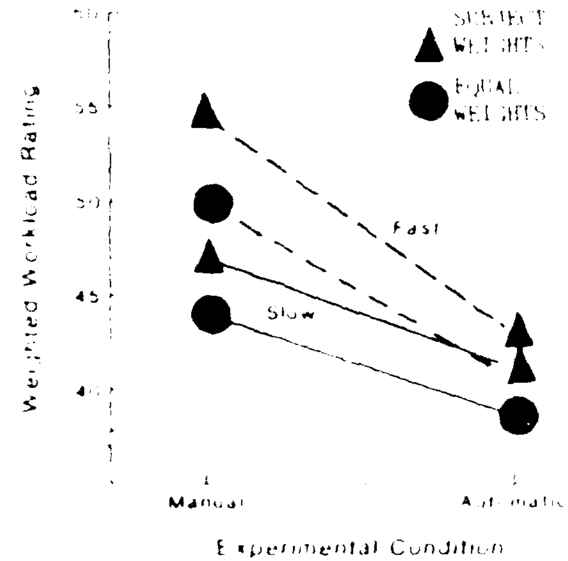


SUBJECTIVE RATINGS: NASA-TASK LOAD INDEX (TLX)

EQUALLY WEIGHTED SUBSCALE RATINGS WERE COMPARED TO RATINGS COMBINED ACCORDING TO THE SUBJECTIVE IMPORTANCE OF INDIVIDUAL FACTORS TO EACH SUBJECT FOR MANUAL AND AUTOMATIC VERSIONS OF A SCHEDULING TASK.

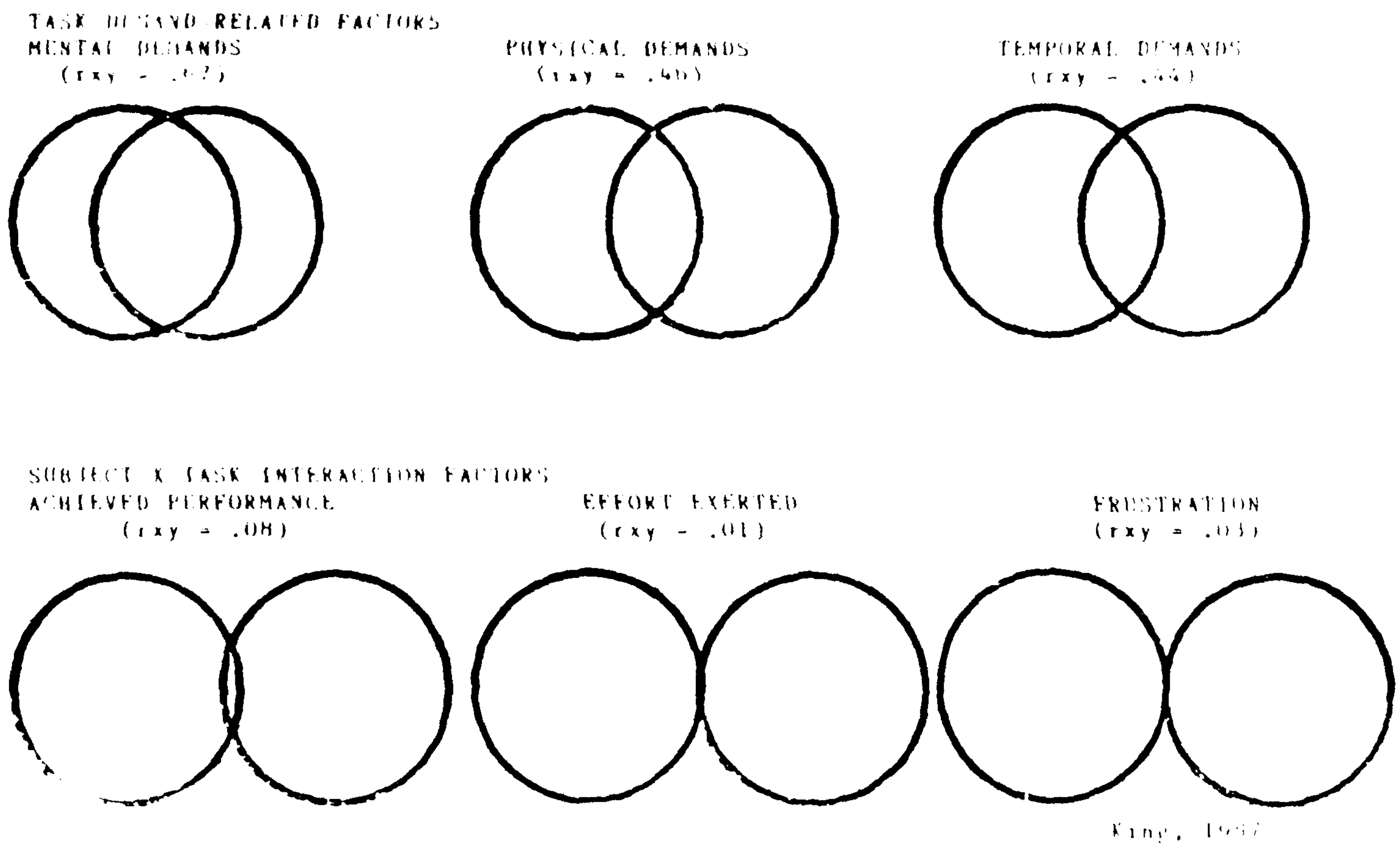
- TLX RATINGS WERE SENSITIVE TO OPERATOR RESPONSIBILITY, BUT NOT TO RATE OF PRESENTATION (IN AGREEMENT WITH POPCORN TASK RESULTS)
- SUBJECT WEIGHTS PROVIDED A MORE SENSITIVE MEASURE THAN EQUAL WEIGHTS
- IMPORTANCE WEIGHTS AND MAGNITUDE RATINGS PROVIDED INDEPENDENT INFORMATION ABOUT THE STRUCTURE OF THE TASK





SUBJECTIVE RATINGS: NASA-TASK LOAD INDEX (NASA-TLX)
CORRELATIONS BETWEEN SUBSCALE RATINGS AND TASK RELATED WEIGHTS

IN SOME CASES THE IMPORTANCE ASSIGNED TO A FACTOR (ITS WEIGHT) MAY COVARY WITH THE MAGNITUDES OF THE RATINGS GIVEN TO THAT SUBSCALE.

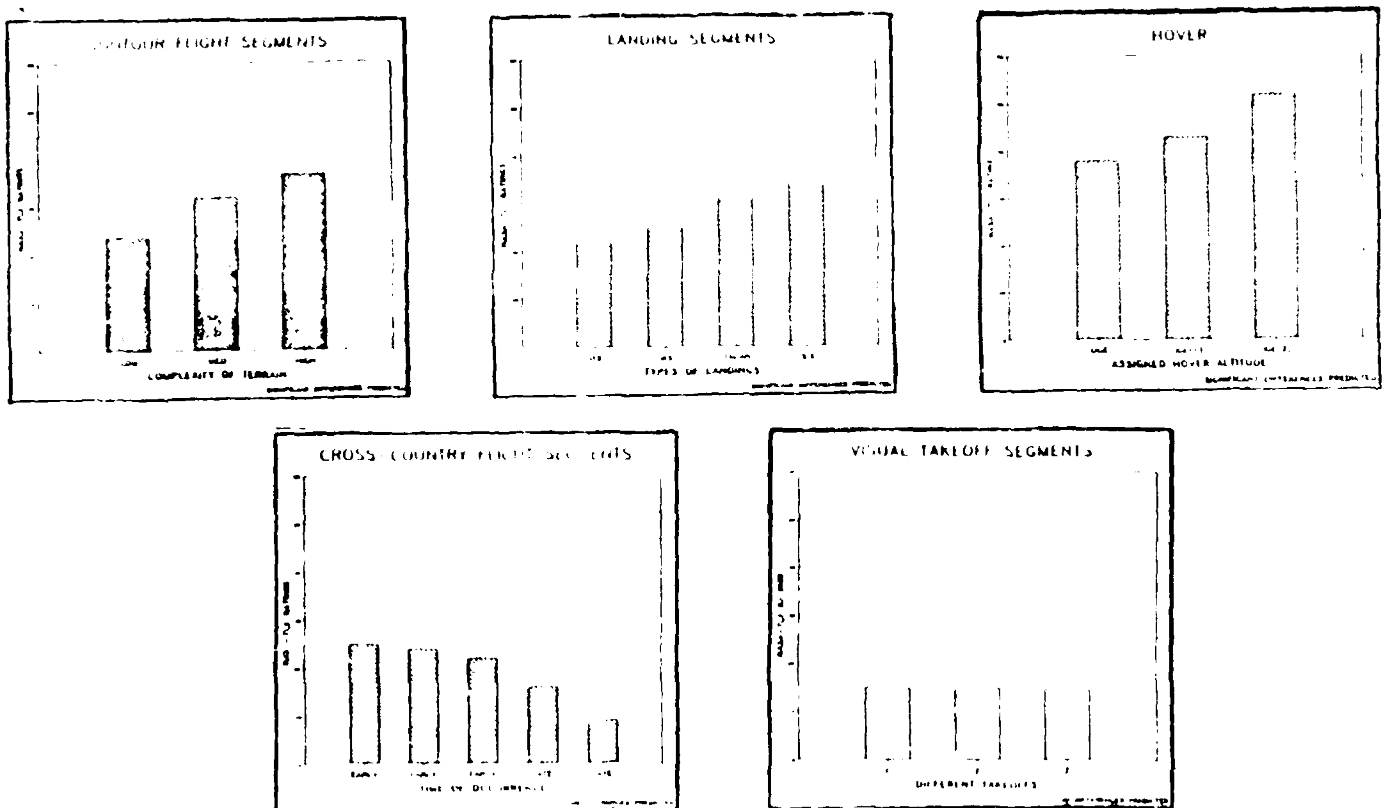


King, 1987

SUBJECTIVE RATINGS: NASA-TASK LOAD INDEX (NASA-TLX) EVIDENCE OF SENSITIVITY

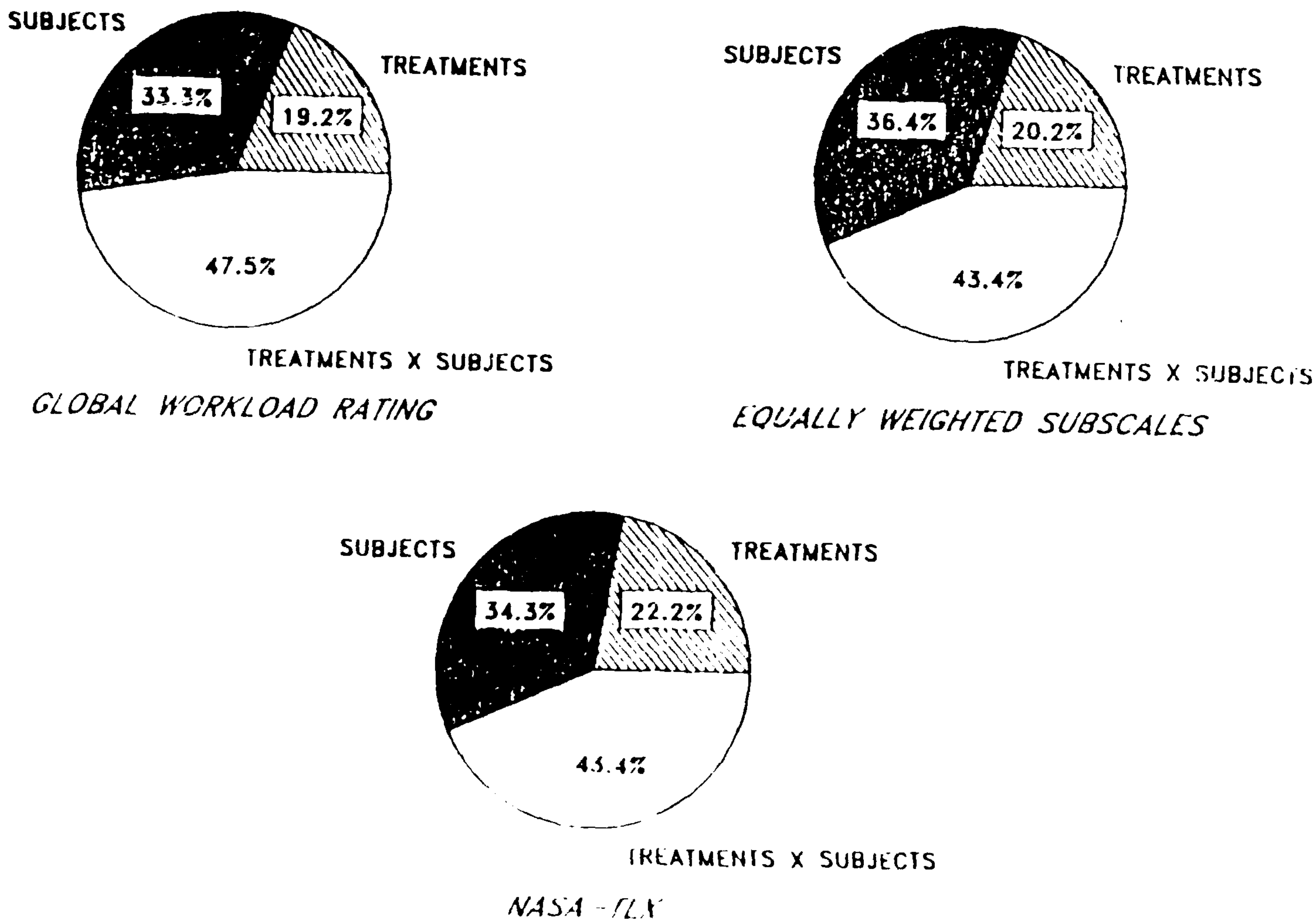
RATING SCALES MUST DEMONSTRATE SENSITIVITY WHERE WORKLOAD DIFFERENCES DO EXIST AND NO SENSITIVITY WHERE THEY DO NOT.

EXAMPLE: SIGNIFICANT DIFFERENCES AMONG NASA TLX RATINGS WERE FOUND AMONG SOME FLIGHT SEGMENTS (AND NOT OTHERS), AS PREDICTED, IN AN SH 3G HELICOPTER.



Shively et al., in press

COMPARISONS AMONG ESTIMATES OF WORKLOAD

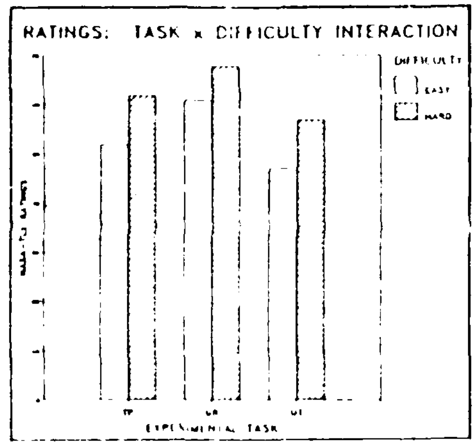


SUBJECTIVE RATINGS: NASA-TASK LOAD INDEX (NASA-TLX) DIAGNOSTICITY

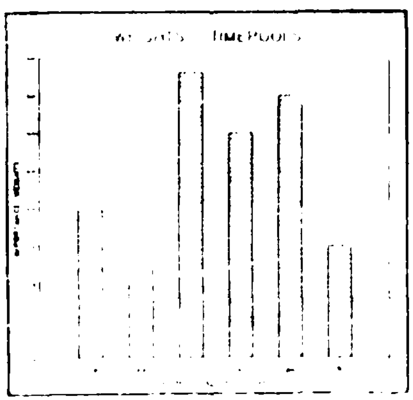
GENERALLY, IT IS NOT SUFFICIENT FOR RATING SCALES TO PROVIDE INFORMATION ABOUT THE MAGNITUDES OF TASK DEMANDS. THEY MUST REFLECT THEIR SOURCE(S) AS WELL.

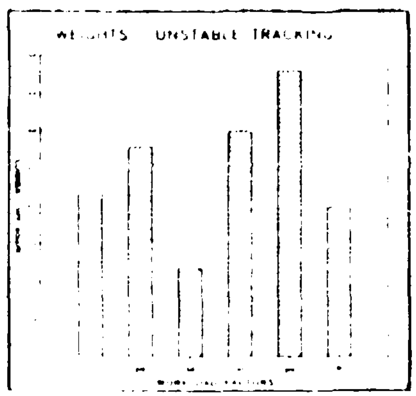
SUBJECTS PROVIDED TLX RATINGS AND IMPORTANCE WEIGHTS FOLLOWING EACH OF THREE TYPES OF TASKS

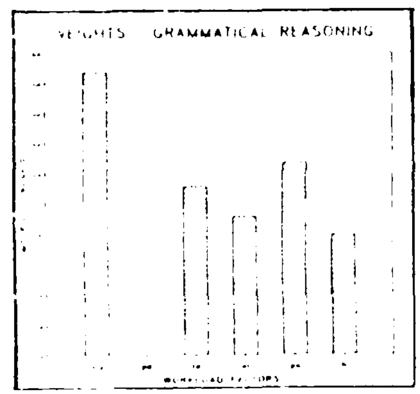
- TIMEPOOLS TARGET ACQUISITION
- GRAMMATICAL REASONING
- UNSTABLE TRACKING



SIGNIFICANT DIFFERENCES WERE FOUND IN THE SUBJECTIVE CONTRIBUTION OF DIFFERENT FACTORS TO THE WORKLOAD OF EACH TASK, PROVIDING ADDITIONAL DIAGNOSTIC INFORMATION







King, 1987

CONCLUSIONS

- WORKLOAD IS A MULTI-DIMENSIONAL CONSTRUCT

- SUBJECTIVE RATINGS DO NOT REPRESENT THE INHERENT PROPERTIES OF A TASK BUT EMERGE FROM THE INTERACTION BETWEEN AN OPERATOR, A TASK, AND THE ENVIRONMENT

- TASK-RELATED DIFFERENCES IN SOURCES OF WORKLOAD ARE BETTER PREDICTORS OF WORKLOAD EXPERIENCES THAN A PRIORI SUBJECTIVE BIASES

- A MULTI-DIMENSIONAL EVALUATION PROCEDURE THAT REFLECTS: THE IMPORTANCE OF DIFFERENT FACTORS TO THE WORKLOAD OF A SPECIFIC TASK (THE WEIGHTS) AND THEIR MAGNITUDES (THE RATINGS) PROVIDE:

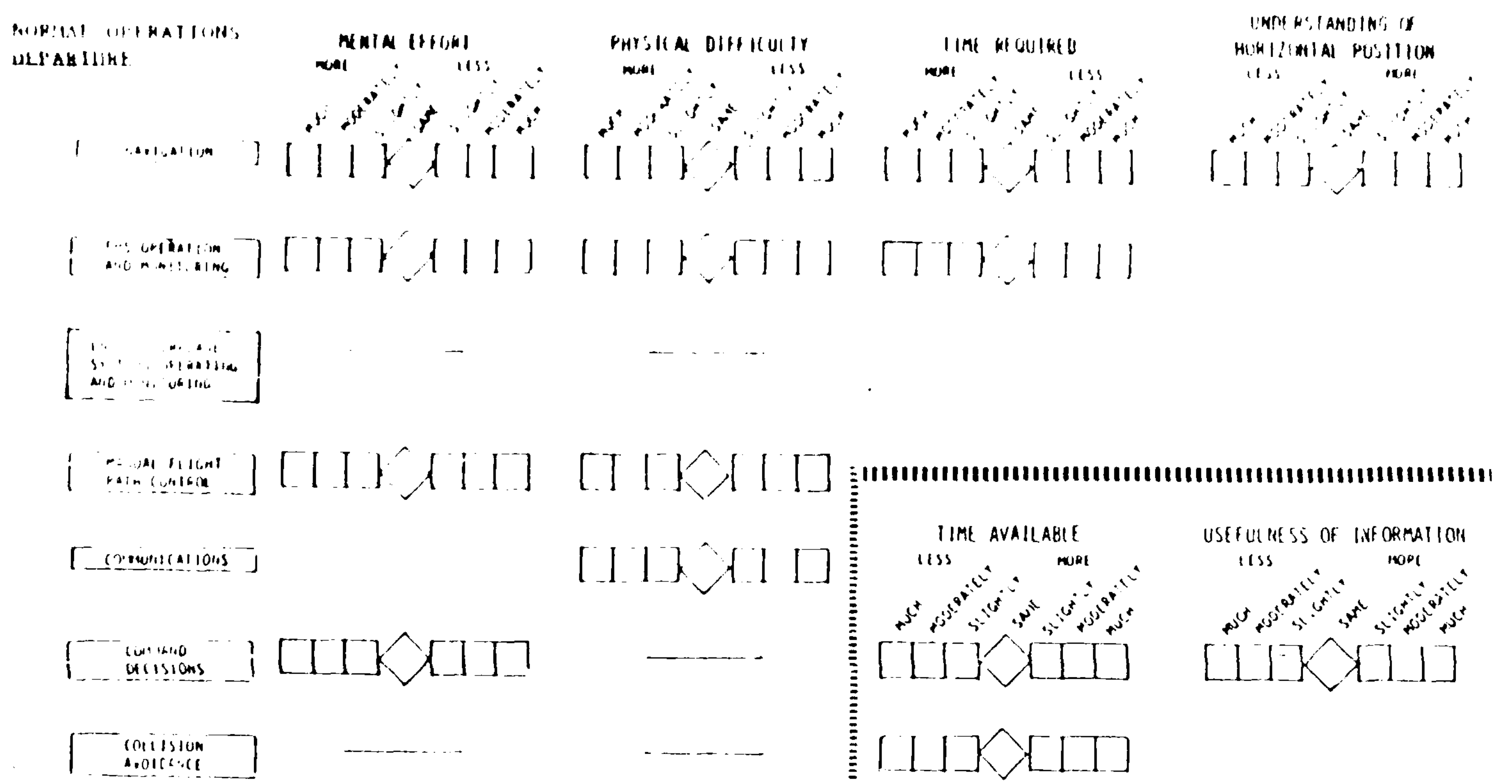
  A SENSITIVE MEASURE OF OVERALL WORKLOAD

  A REDUCTION IN BETWEEN-SUBJECT VARIABILITY

  DIAGNOSTIC INFORMATION ABOUT THE WORKLOAD-STRUCTURE OF A TASK

BOEING AIRCRAFT COMPANY RATING SCALE

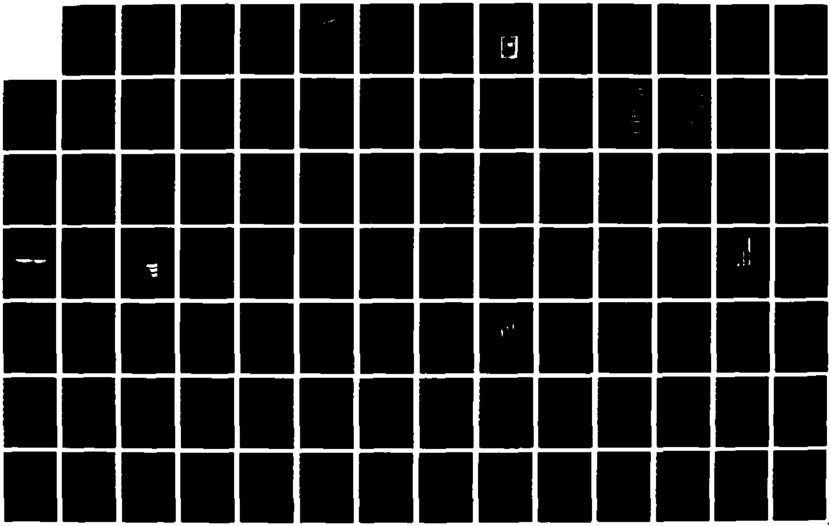
SUBJECTIVE RATINGS: PILOT SUBJECTIVE EVALUATION SCALE

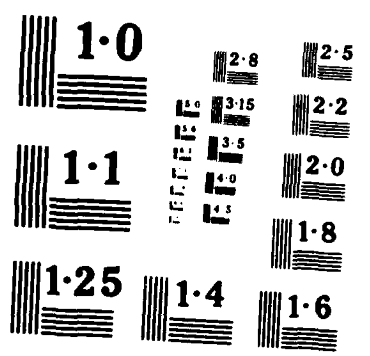


Ruggiero & Fadden, 1982

AD-A189 004 PROCEEDINGS OF THE WORKSHOP ON THE ASSESSMENT OF CREW WORKLOAD MEASUREMEN..(U) DOUGLAS AIRCRAFT CO LONG BEACH CA M A BIFERNO ET AL. JUN 87 AFWAL-TR-87-3043-VOL-1 3/4

UNCLASSIFIED F33615-86-C-3600 F/G 5/9 NL




1·0
2·8
2·5
3·15
2·2
1·1
3·5
2·0
4·0
4·5
1·8
1·25
1·4
1·6

SUBJECTIVE RATINGS:
SUMMARY OF REQUIREMENTS FOR CERTIFICATION

- REASONABLE RELIABILITY
- DEMONSTRATED VALIDITY
  - CONSTRUCT
    - TESTED AGAINST KNOWN WORKLOAD LEVELS
    - PARTICULARLY SENSITIVE TO THE "COST" OF TASK PERFORMANCE TO THE PILOT
  - CONVERGING
    - PROVIDES THE SAME (OR BETTER) INFORMATION THAN OTHER VALID MEASURES
  - ACCUMULATED
    - EVIDENCE OBTAINED FROM SEVERAL SOURCES
- SENSITIVE TO AND DIAGNOSTIC ABOUT:
  - PSYCHOLOGICAL VARIABLES
    - MENTAL DEMANDS, PHYSICAL DEMANDS, STRESS
  - TASK-RELATED VARIABLES
    - FAR-25, PART-D
- USED IN CONJUNCTION WITH:
  - REFERENCE TASKS
  - REFERENCE VEHICLE
- APPROPRIATE FOR USE IN FLIGHT

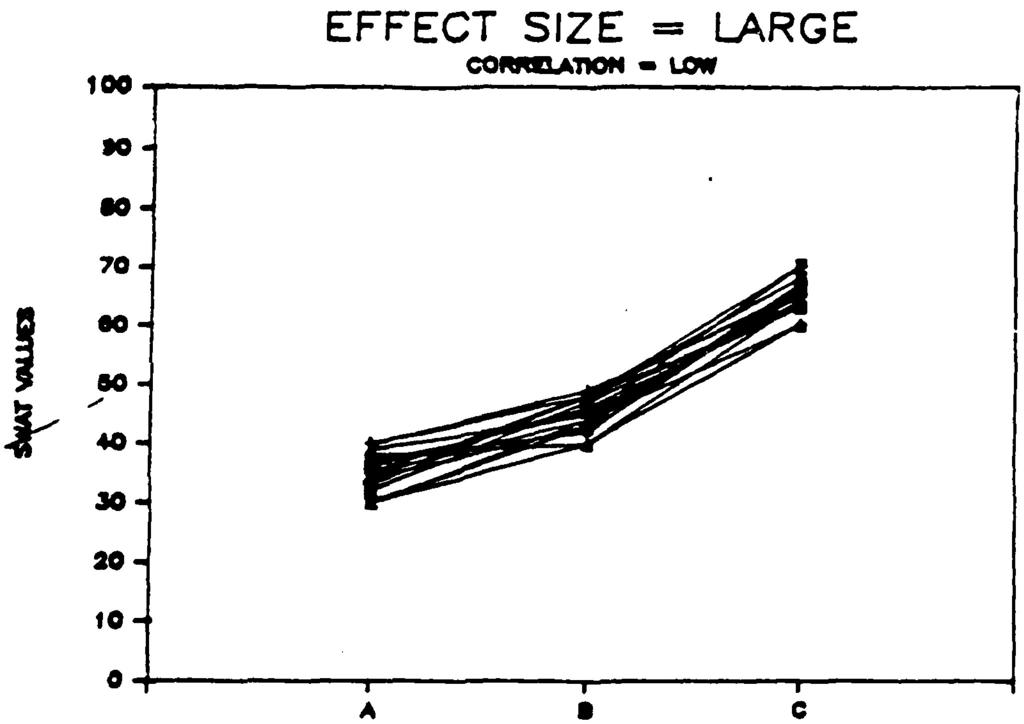
SUMMARY OF WORKLOAD ASSESSMENT TECHNIQUE CAPABILITIES

	SENSITIVITY	DIAGNOSTICITY	INTRUSIVENESS	IMPLEMENTATION REQUIREMENTS	OPERATOR ACCEPTANCE
PRIMARY TASK MEASURES	DISCRIMINATE OVERLOAD FROM NONOVERLOAD SITUATIONS. USED TO DETERMINE IF OPERATOR PERFORMANCE WILL BE ACCEPTABLE WITH A PARTICULAR DESIGN OPTION	NOT CONSIDERED DIAGNOSTIC. REPRESENTS A GLOBAL MEASURE OF WORKLOAD THAT IS SENSITIVE TO OVERLOADS ANYWHERE WITHIN THE OPERATOR'S PROCESSING SYSTEM	NONINTRUSIVE SINCE NO ADDITIONAL OPERATOR PERFORMANCE OR REPORT REQUIRED	INSTRUMENTATION FOR DATA COLLECTION CAN RESTRICT USE IN OPERATIONAL ENVIRONMENTS USE REQUIRES MOCKUPS, SIMULATORS, OR OPERATIONAL EQUIPMENT. IMPOSES LIMITS ON USE DURING EARLY SYSTEM DEVELOPMENT NO OPERATOR TRAINING REQUIRED	NO SYSTEMATIC DATA. NO REASON TO EXPECT NEGATIVE OPERATOR OPINION
SECONDARY TASK METHODS	CAPABLE OF DISCRIMINATING LEVELS OF CAPACITY EXPENDITURE IN NONOVERLOAD SITUATIONS. USED TO ASSESS RESERVE CAPACITY AFFORDED BY A PRIMARY TASK. CAN BE USED TO [illegible] THE POTENTIAL FOR [illegible] AMONG DESIGN OPTIONS	CAPABLE OF DISCRIMINATING SOME DIFFERENCES IN RESOURCE EXPENDITURE (E.G. CENTRAL PROCESSING VERSUS MOTOR). DIAGNOSTICITY SUGGESTS COMPLEMENTARY USE WITH MORE GENERALLY SENSITIVE MEASURES, WITH THE LATTER INITIALLY IDENTIFYING OVERLOADS AND SECONDARY TASKS BEING USED SUBSEQUENTLY TO PINPOINT THE LOCUS OF OVERLOAD	PRIMARY TASK INTRUSION HAS REPRESENTED A PROBLEM IN MANY APPLICATIONS, PARTICULARLY IN THE LABORATORY. DATA ARE NOT EXTENSIVE IN OPERATIONAL ENVIRONMENTS. SEVERAL TECHNIQUES (E.G. EMBEDDED SECONDARY TASK, ADAPTIVE PROCEDURES) HAVE BEEN DESIGNED TO CONTROL INTRUSION. POTENTIAL FOR INTRUSION COULD LIMIT USE IN OPERATIONAL ENVIRONMENTS	INSTRUMENTATION FOR DATA COLLECTION CAN RESTRICT USE IN OPERATIONAL ENVIRONMENT, BUT SOME TASKS HAVE BEEN INSTRUMENTED FOR IN-FLIGHT USE. USE REQUIRES MOCKUPS, SIMULATORS, OR OPERATIONAL EQUIPMENT. IMPOSES LIMITS ON USE DURING EARLY SYSTEM DEVELOPMENT SOME OPERATOR TRAINING USUALLY REQUIRED TO STABILIZE SECONDARY TASK PERFORMANCE	NO SYSTEMATIC DATA. REQUIREMENT TO PERFORM SECONDARY TASK COULD DISTRACT OPERATOR. TECHNIQUE SUCH AS EMBEDDED SECONDARY TASK SHOULD MINIMIZE ANY ACCEPTANCE PROBLEMS
PHYSIOLOGICAL TECHNIQUES	CAPABLE OF DISCRIMINATING LEVELS OF CAPACITY EXPENDITURE IN NONOVERLOAD SITUATIONS. CAN BE USED TO ASSESS THE RELATIVE POTENTIAL FOR OVERLOAD AMONG DESIGN OPTIONS	SOME TECHNIQUES (E.G. EVENT RELATED BRAIN POTENTIAL) APPEAR DIAGNOSTIC OF SOME RESOURCES, WHILE OTHER MEASURES (E.G. PUPIL DIAMETER) APPEAR MORE GENERALLY SENSITIVE. CHOICE OF TECHNIQUE DEPENDENT ON PURPOSE OF MEASUREMENT (SCREENING FOR ANY OVERLOAD VERSUS IDENTIFYING LOCUS OF OVERLOAD)	INTRUSION DOES NOT APPEAR TO REPRESENT A MAJOR PROBLEM ALTHOUGH THERE ARE DATA TO INDICATE THAT SOME INTERFERENCE CAN OCCUR	INSTRUMENTATION FOR DATA COLLECTION CAN RESTRICT USE IN OPERATIONAL ENVIRONMENTS. USE REQUIRES MOCKUPS, SIMULATORS, OR OPERATIONAL EQUIPMENT. IMPOSES LIMITS ON USE DURING EARLY SYSTEM DEVELOPMENT NO OPERATOR TRAINING REQUIRED	NO SYSTEMATIC DATA. INSTRUMENTATION AND RECORDING EQUIPMENT COULD REPRESENT POTENTIAL PROBLEM, BUT NO SIGNIFICANT PROBLEMS REPORTED IN LITERATURE
SUBJECTIVE TECHNIQUES	CAPABLE OF DISCRIMINATING LEVELS OF CAPACITY EXPENDITURE IN NONOVERLOAD SITUATIONS. CAN BE USED TO ASSESS THE RELATIVE POTENTIAL FOR OVERLOAD AMONG DESIGN OPTIONS	NOT CONSIDERED DIAGNOSTIC. AVAILABLE EVIDENCE INDICATES THAT RATING SCALES REPRESENT A GLOBAL MEASURE OF LOAD. LACK OF DIAGNOSTICITY SUGGESTS USE AS A GENERAL SCREENING DEVICE TO DETERMINE IF OVERLOAD EXISTS ANYWHERE WITHIN TASK PERFORMANCE	INTRUSION DOES NOT APPEAR TO REPRESENT A SIGNIFICANT PROBLEM. MOST APPLICATIONS REQUIRE RATING SCALE COMPLETION SUBSEQUENT TO TASK PERFORMANCE AND THEREFORE PRESENT NO INTRUSION PROBLEM	INSTRUMENTATION REQUIRED IS USUALLY MINIMAL, PERMITTING USE IN A NUMBER OF ENVIRONMENTS. TRADITIONAL APPLICATIONS REQUIRE MOCKUPS, SIMULATORS, OR OPERATIONAL EQUIPMENT. IMPOSES LIMITS ON USE DURING EARLY SYSTEM DEVELOPMENT. RECENT PROJECTIVE USE PROVIDES POTENTIAL FOR APPLICATION DURING EARLY STAGES. SOME FAMILIARIZATION WITH PROCEDURES CAN BE REQUIRED	NO SYSTEMATIC DATA. INFORMAL EVIDENCE SUGGESTS THAT SEVERAL RATING SCALES ENJOY A HIGH DEGREE OF OPERATOR ACCEPTANCE

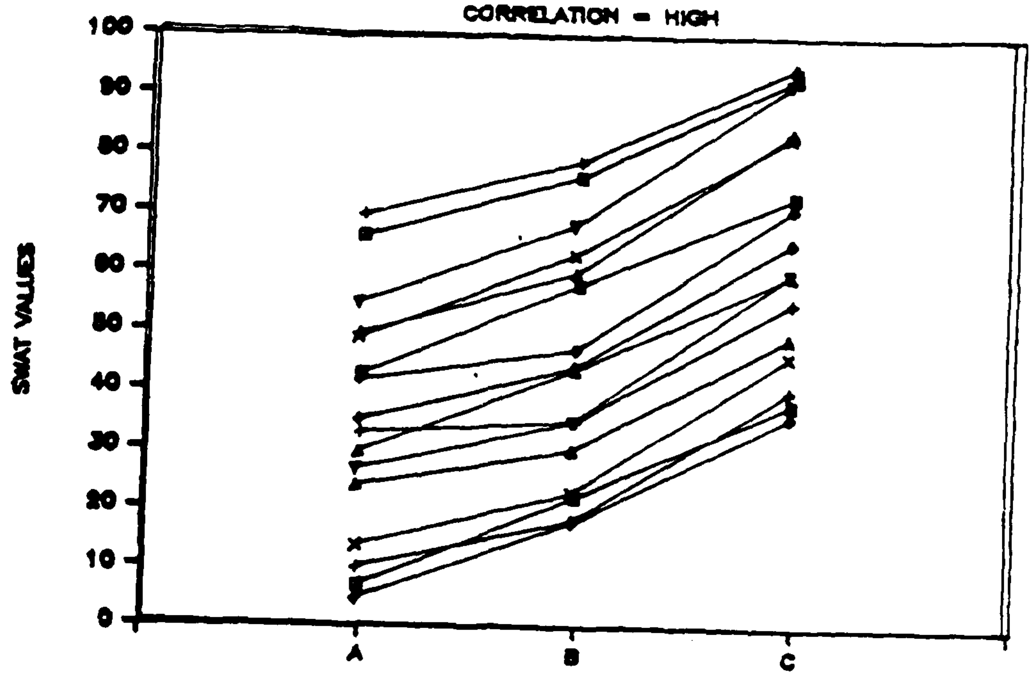
SUBJECTIVE RATINGS: RELIABILITY

- PARTICULARLY FOR AIRCRAFT CERTIFICATION, RELIABILITY IS AN IMPORTANT CONSIDERATION BECAUSE:
  - o THE NUMBER OF EVALUATION PILOTS MAY BE LIMITED
  - o REPEATED MEASUREMENTS FOR THE SAME PILOT ARE COSTLY
  - o THE CONSEQUENCES OF AMIBGUOUS OR INACCURATE RESULTS ARE UNACCEPTABLE

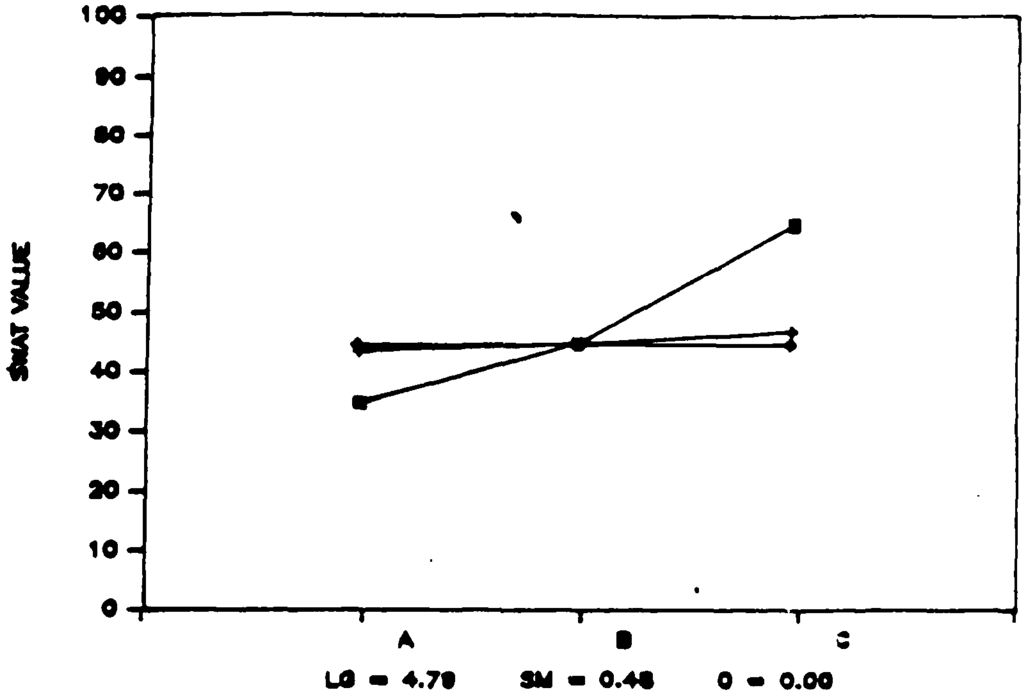
- METHODS OF EVALUATING RELIABILITY:
  - o SPLIT-HALF
  - o TEST-RETEST
  - o INTER-RATER
  - o ALTERNATE FORMS


EFFECT SIZE = LARGE
CORRELATION = LOW
SWAT VALUES
100
90
80
70
60
50
40
30
20
10
0
A
B
C

EFFECT SIZE = LARGE

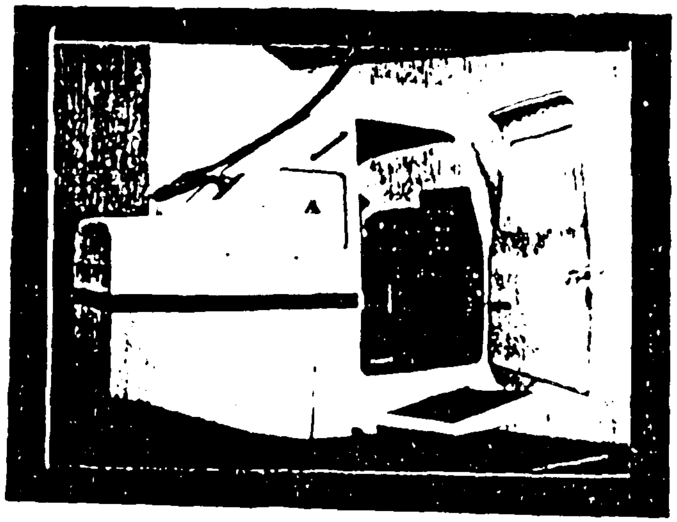


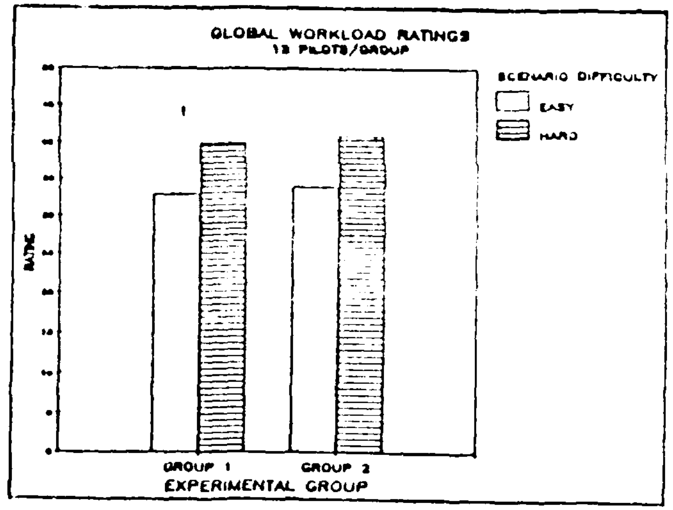
EFFECT SIZE



SUBJECTIVE RATING: SPLIT-HALF RELIABILITY

EXAMPLE: RATINGS WERE OBTAINED FROM TWO DIFFERENT GROUPS OF 12 PILOTS AFTER PERFORMING AN 'EASY' AND A 'HARD' FLIGHT IN A MOTION-BASE SIMULATOR USING THE NASA BIPOLAR RATING SCALE.



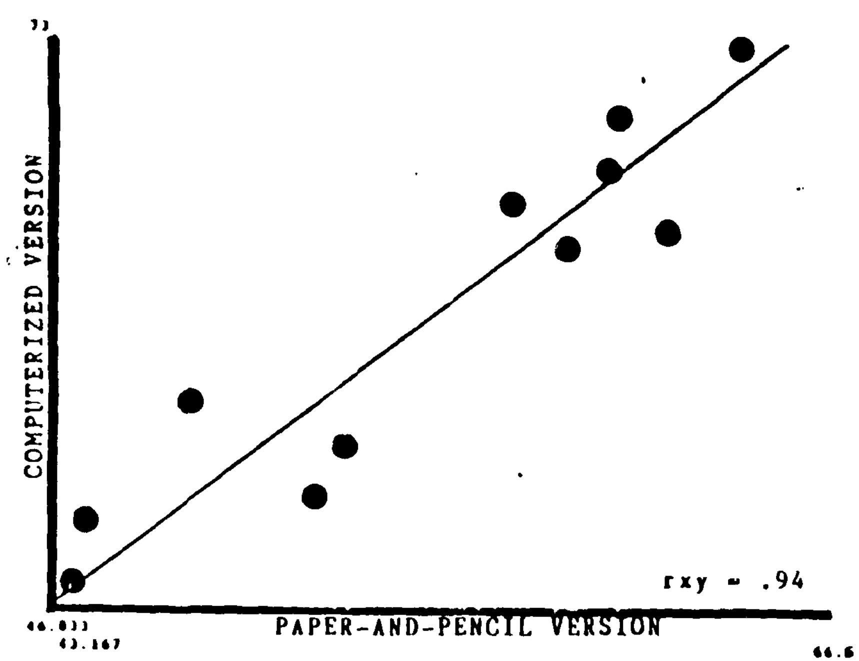


SUBJECTIVE RATINGS: ALTERNATE FORMS RELIABILITY

EXAMPLE: RATINGS WERE OBTAINED FROM SIX SUBJECTS USING THREE FORMS OF THE NASA-TASK LOAD INDEX: COMPUTERIZED, PAPER-AND-PENCIL, AND VERBAL

TASKS: TARGET ACQUISITION, GRAMMATICAL REASONING, UNSTABLE TRACKING

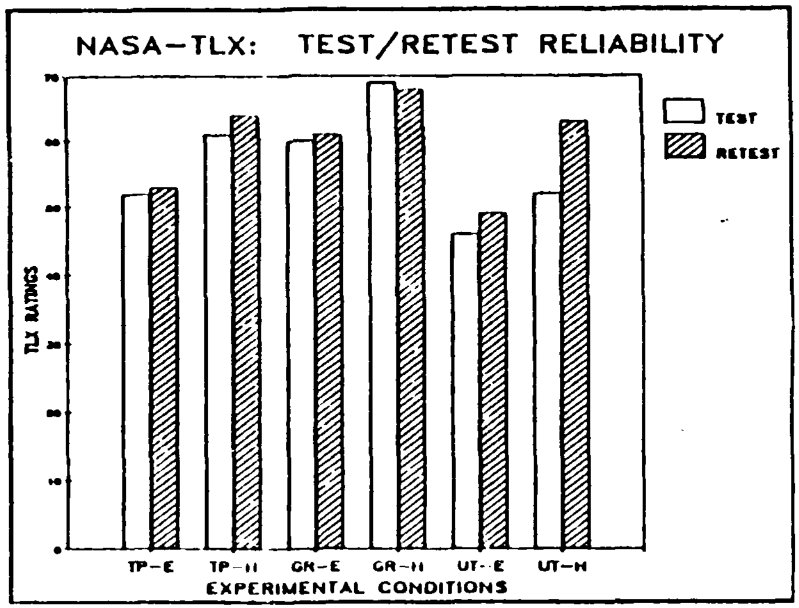
RESULTS: RATINGS WITH ALTERNATE FORMS OF NASA-TLX WERE HIGHLY CORRELATED



King, 1987

SUBJECTIVE RATINGS: NASA-TASK LOAD INDEX (NASA-TLX) TEST/RETEST RELIABILITY

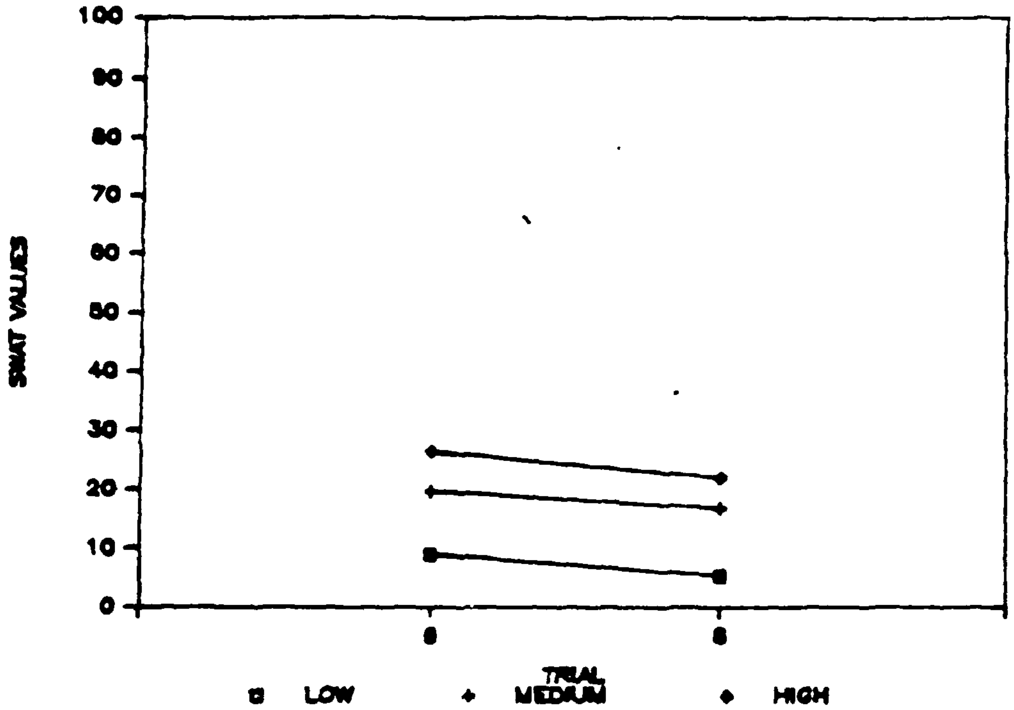
SUBJECTS PROVIDED TLX RATINGS FOLLOWING ASYMPTOTIC PERFORMANCE OF TWO LEVELS (E,H) OF THREE TASKS (TIMEPOOLS TARGET ACQUISITION, GRAMMATICAL REASONING, AND UNSTABLE TRACKING) TWICE, SEPARATED BY AN INTERVAL OF 4 WEEKS. NO SIGNIFICANT DIFFERENCES WERE FOUND.



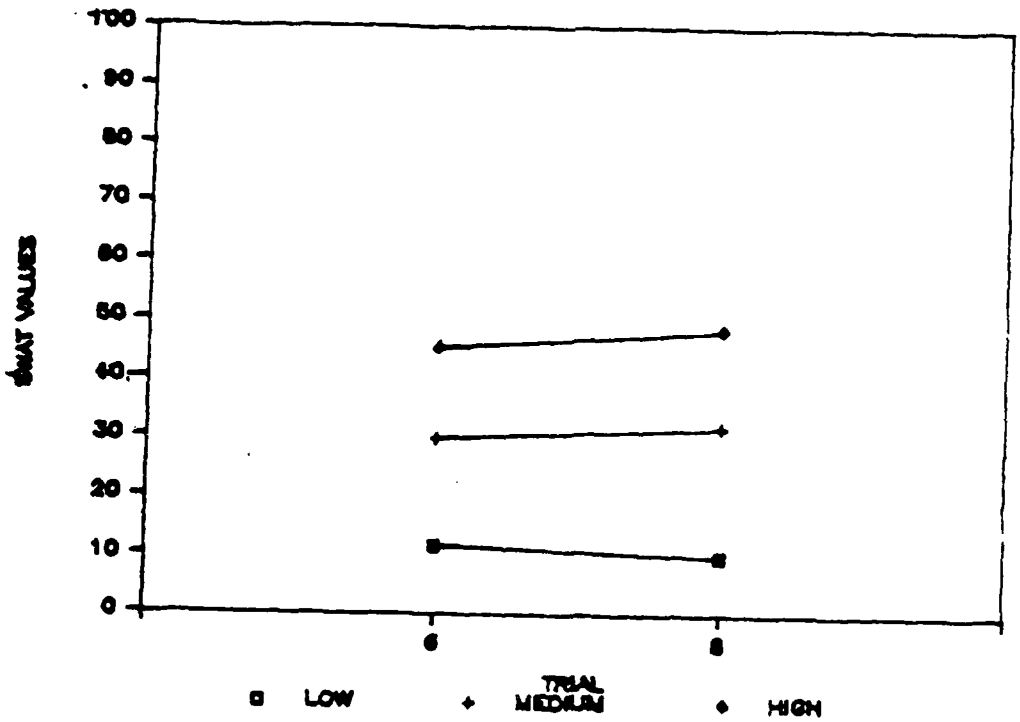
NOTE: CORRELATION IS .83

King, 1987

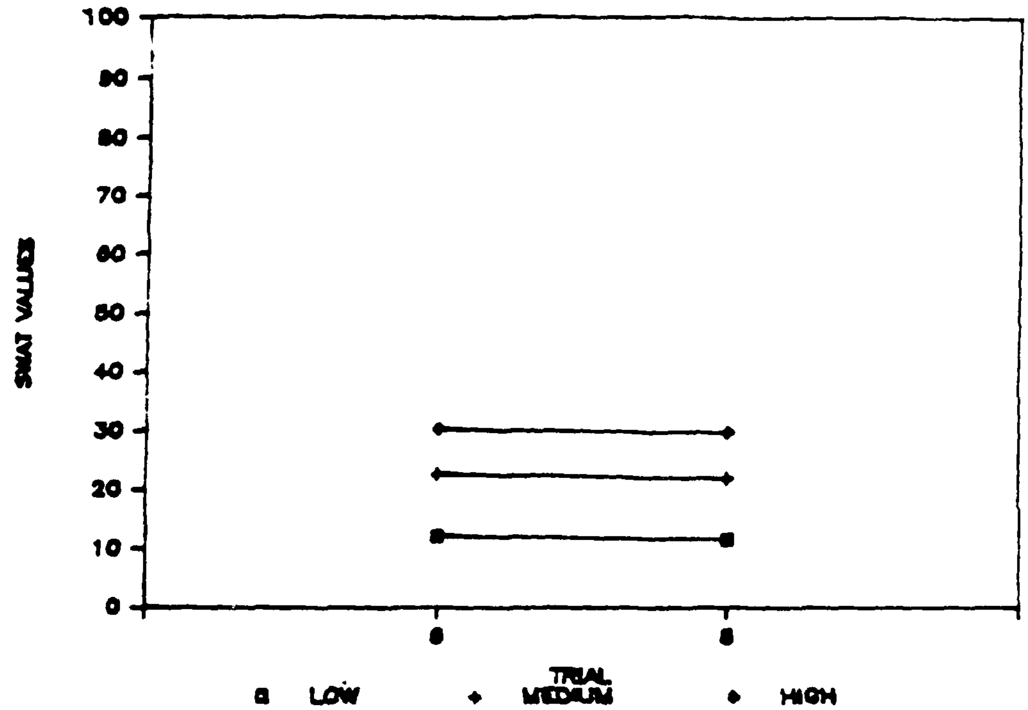
SPATIAL PROCESSING

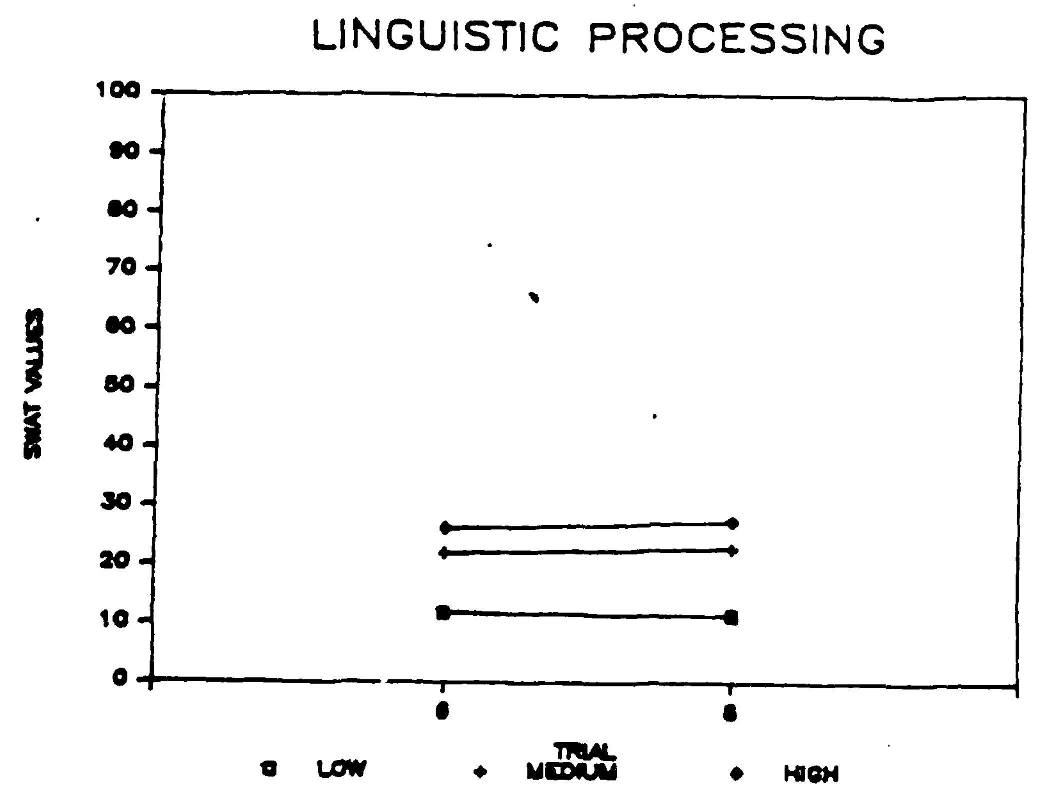


PROBABILITY MONITORING

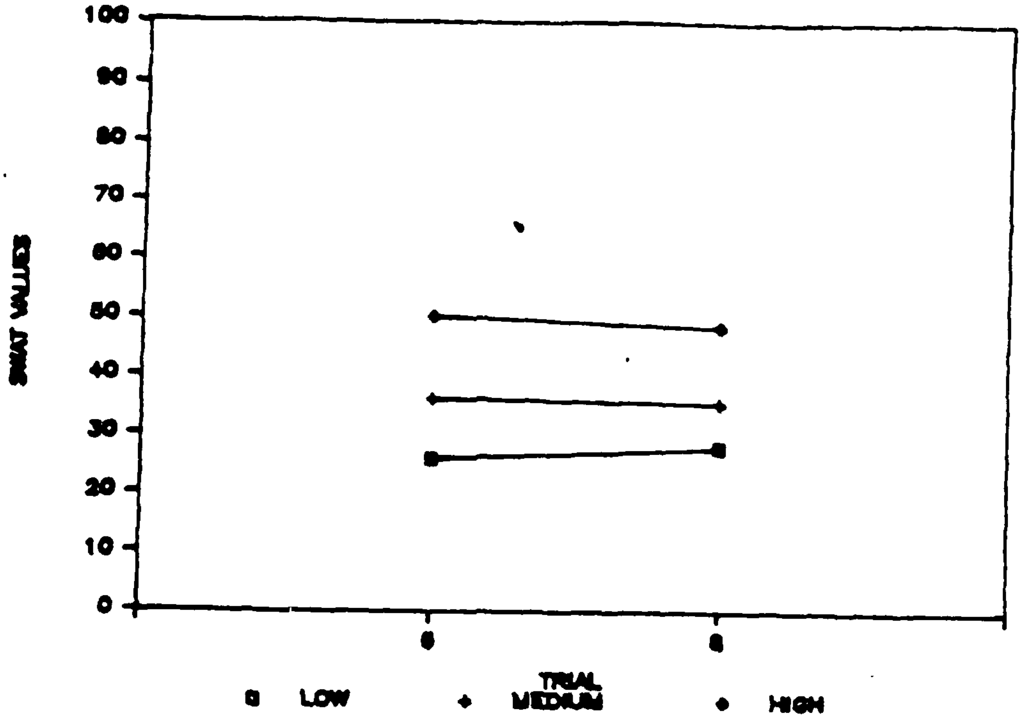


MATH PROCESSING

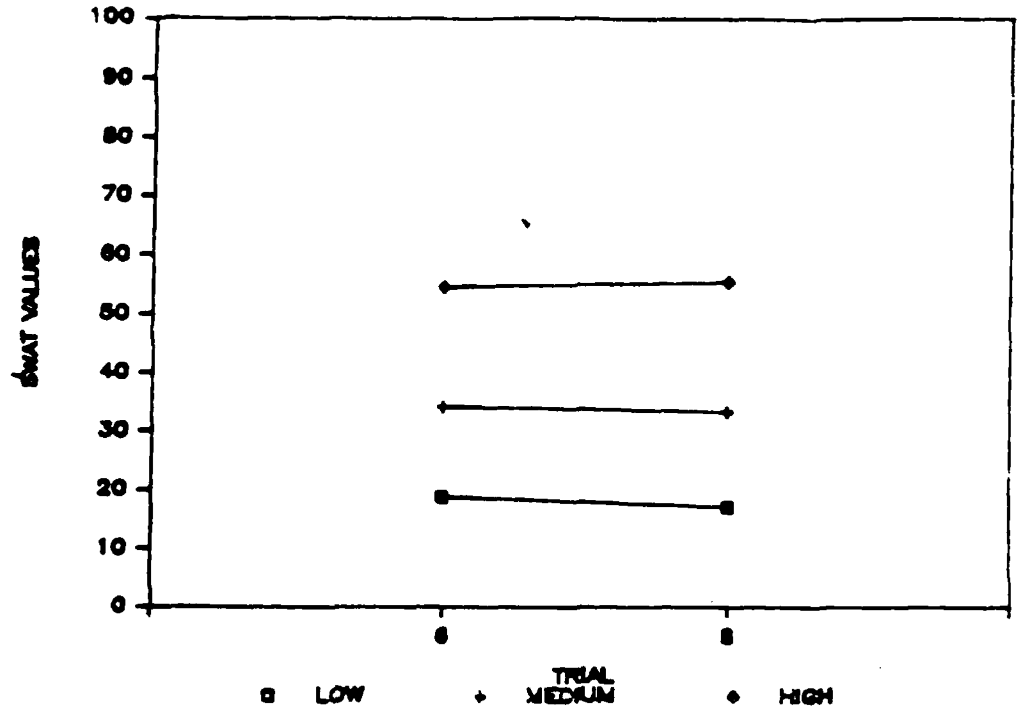



LINGUISTIC PROCESSING
100
90
80
70
60
50
40
30
20
10
0
SWAT VALUES
TRIAL
LOW
MEDIUM
HIGH

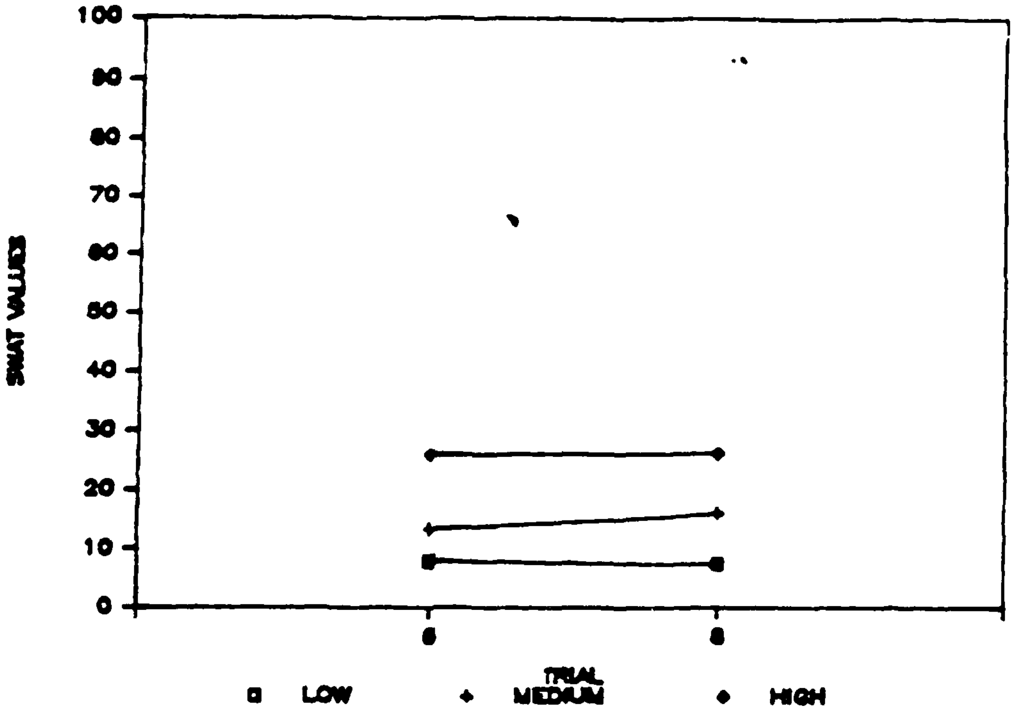
GRAMMATICAL REASONING



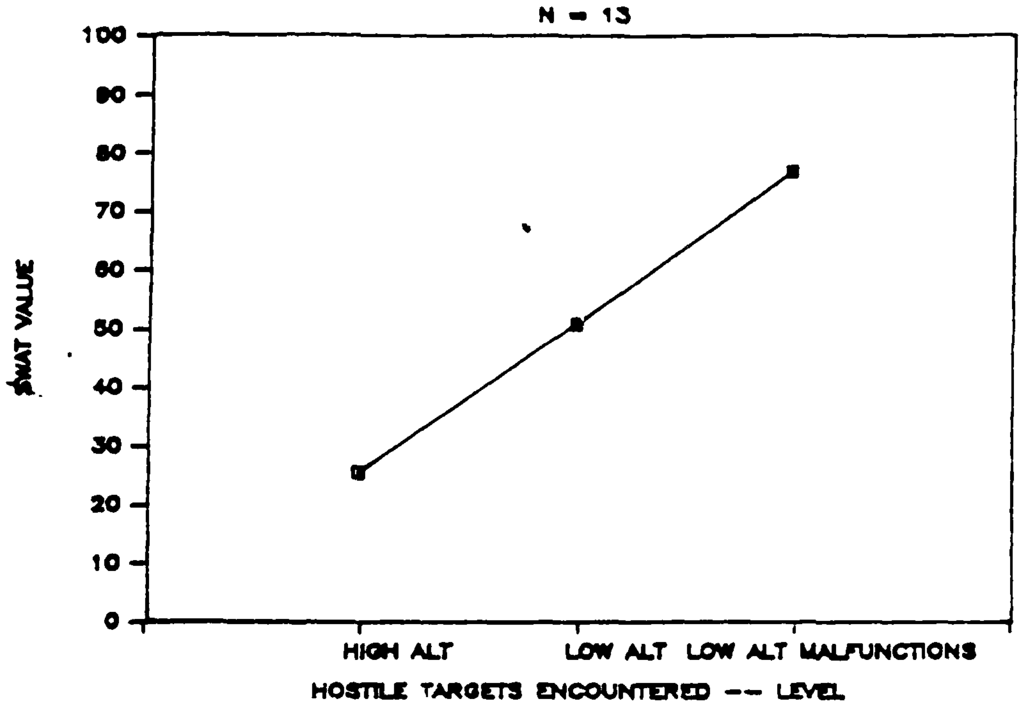
CONTINUOUS RECALL

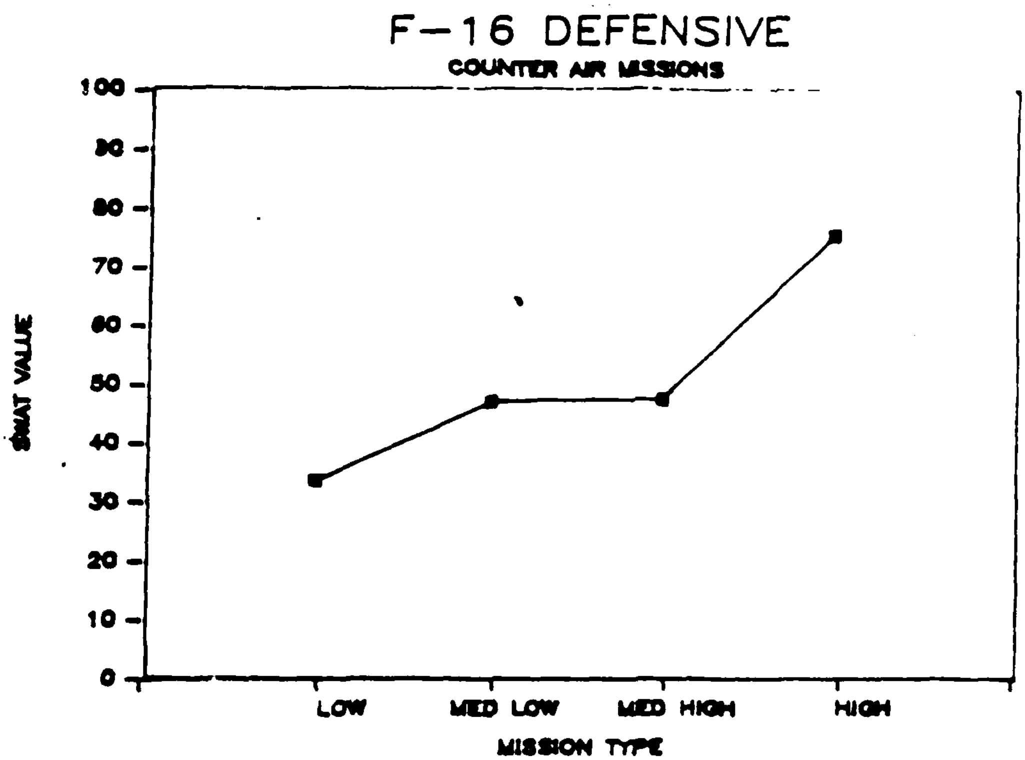


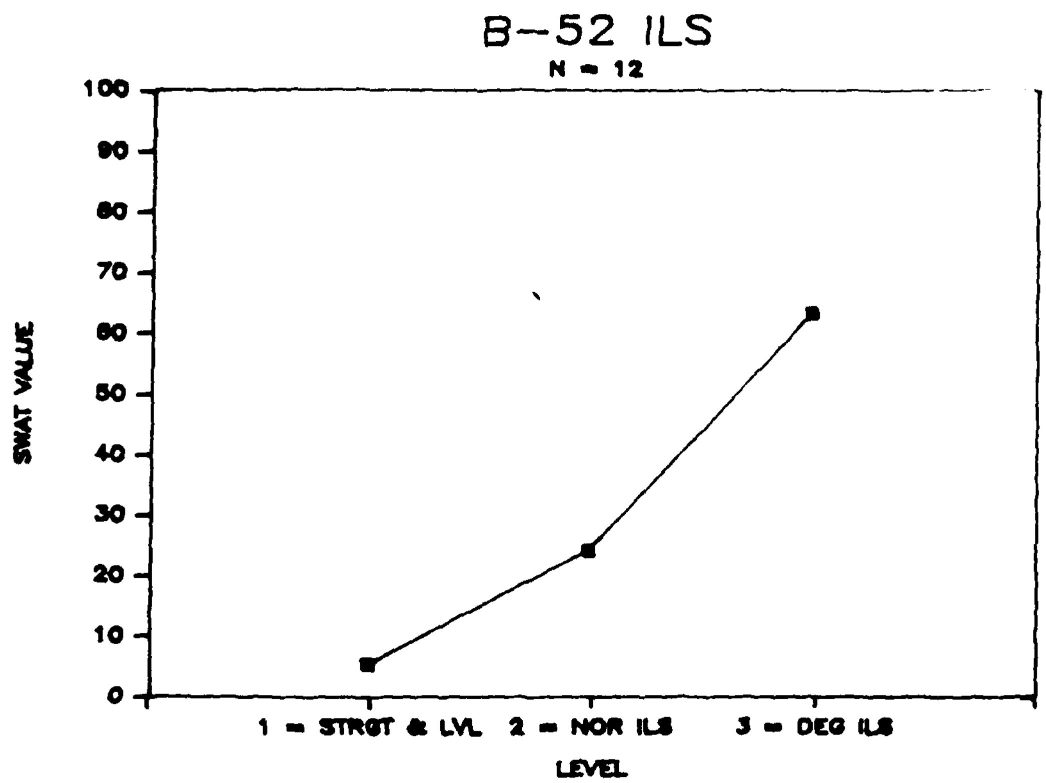
MEMORY SEARCH



B-52 TAIL GUNNER




F-16 DEFENSIVE
COUNTER AIR MISSIONS
SWAT VALUE
100
90
80
70
60
50
40
30
20
10
0
LOW
MED LOW
MED HIGH
HIGH
MISSION TYPE


B-52 ILS
N = 12
100
90
80
70
60
50
40
30
20
10
0
SWAT VALUE
1 = STRGT & LVL
2 = NOR ILS
3 = DEG ILS
LEVEL

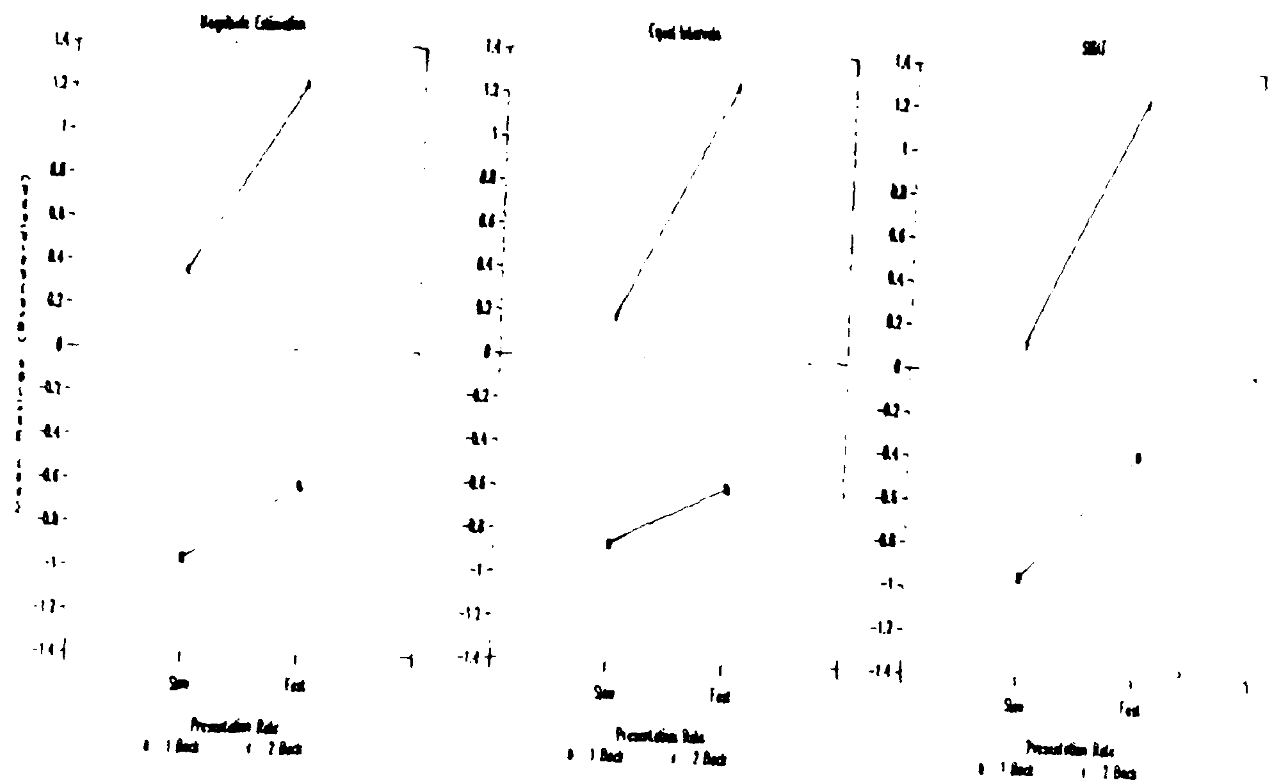


Figure 2: Number back by presentation rate interaction for each technique.

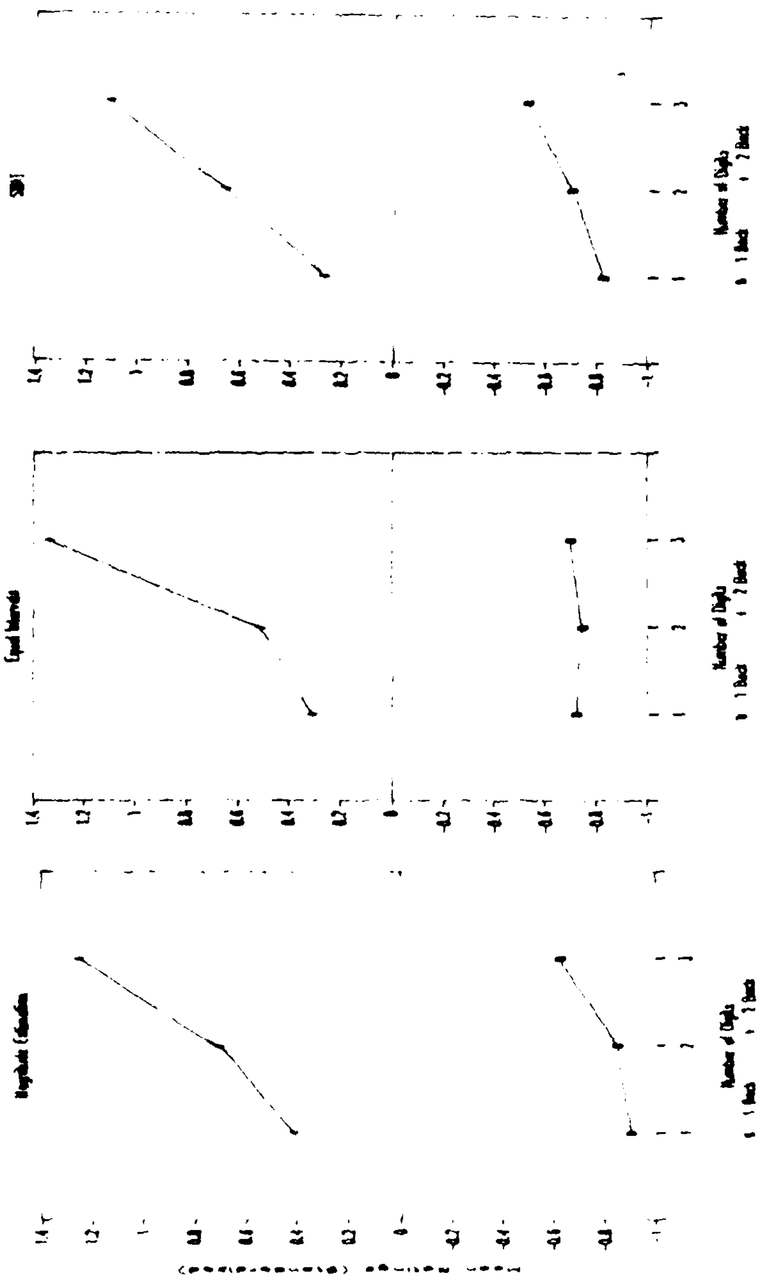


Figure 3: Number back by number of digits interaction for each technique.

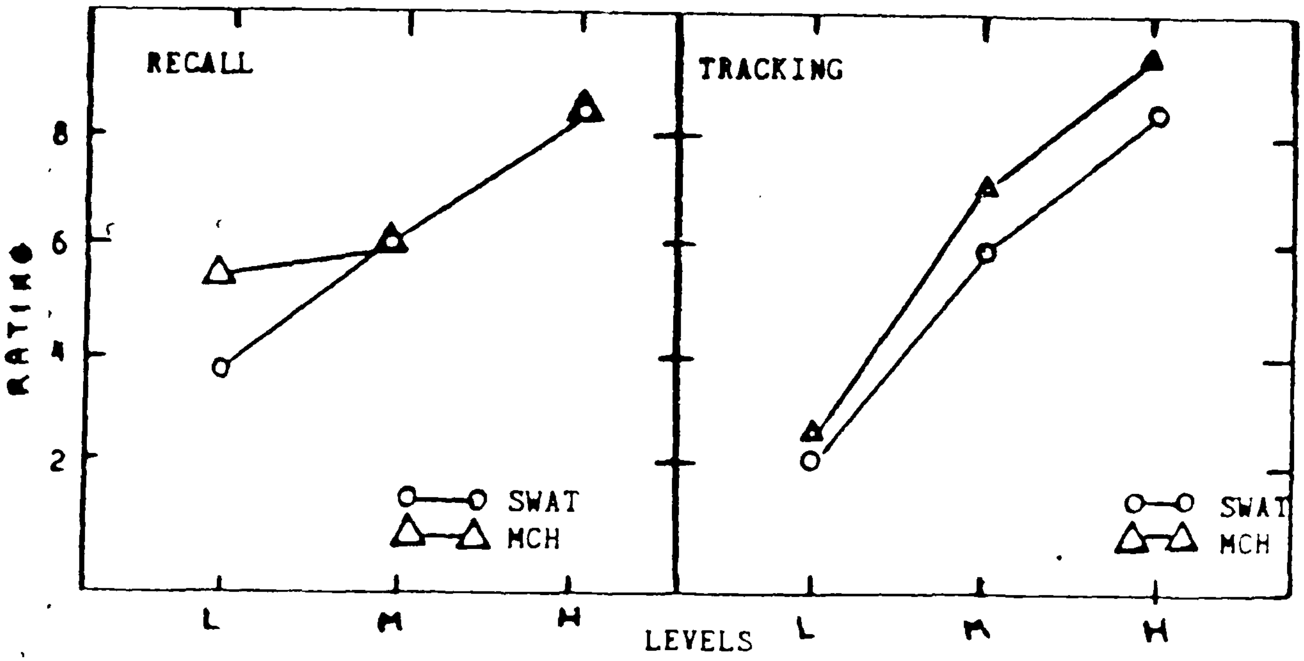
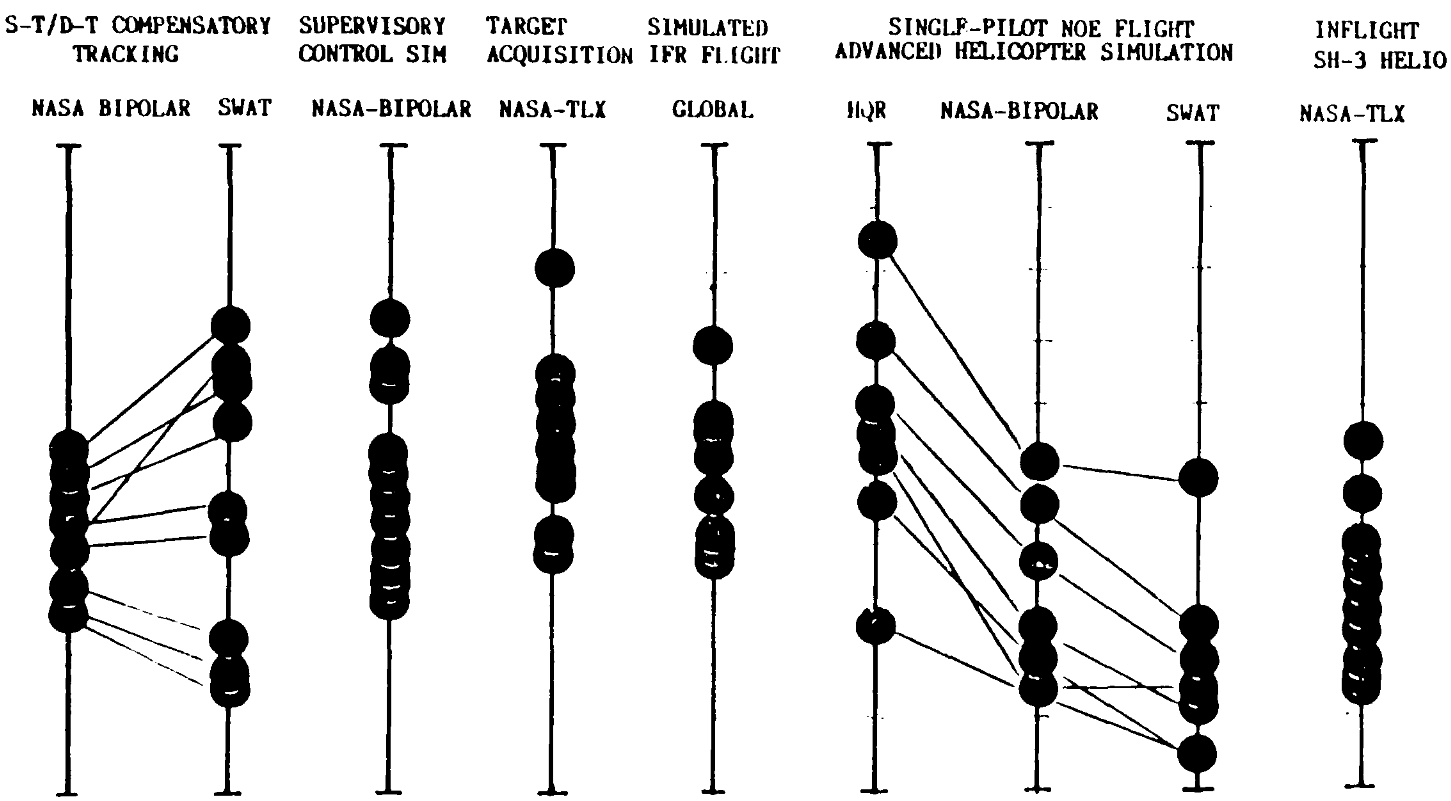


Figure 1. Subjective ratings as a function of task difficulty manipulations on two types of tasks.

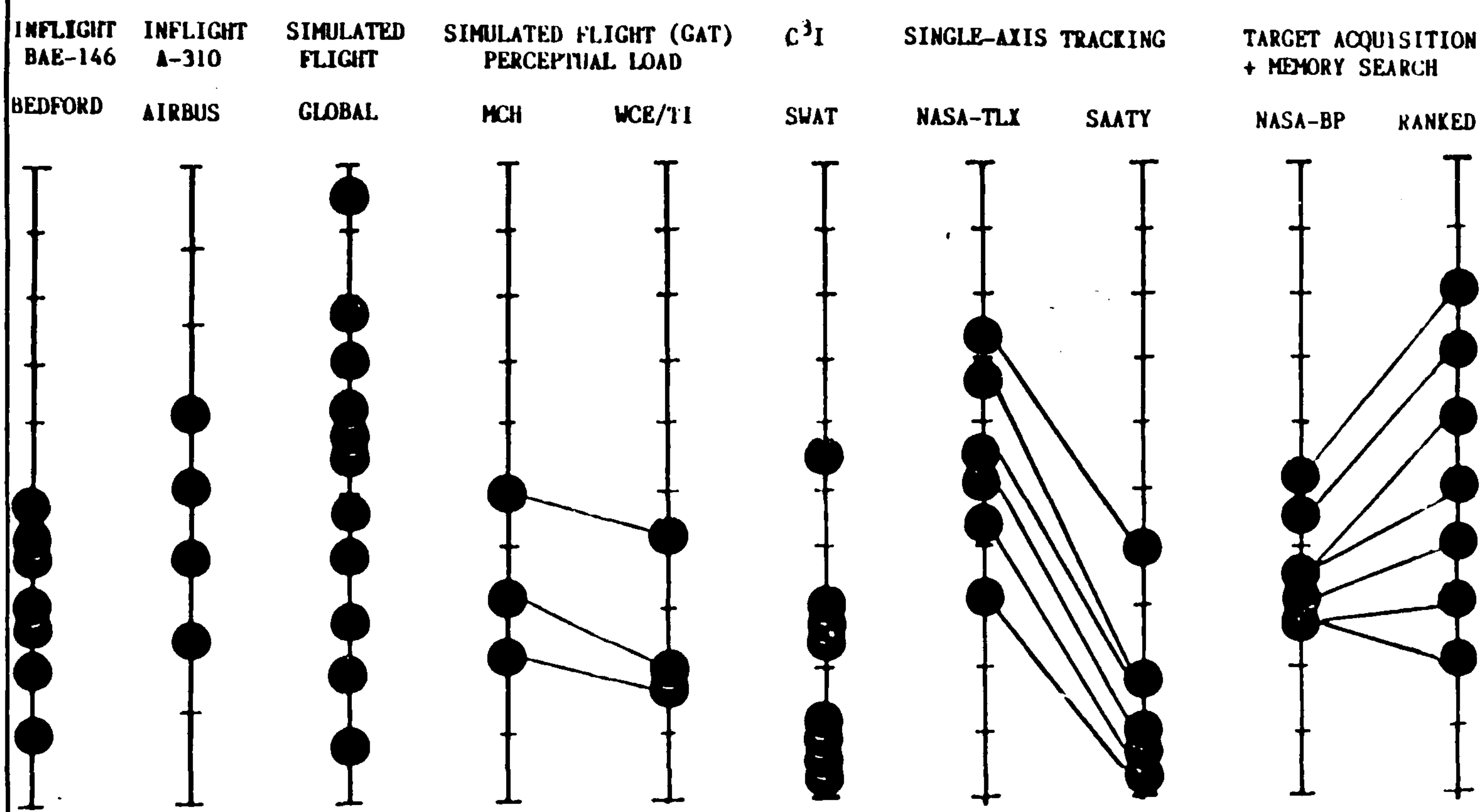
SUBJECTIVE RATINGS: MEASUREMENT ERROR CONTEXT EFFECTS

AVERAGE WORKLOAD RATINGS OBTAINED IN DIFFERENT EXPERIMENTS USING:

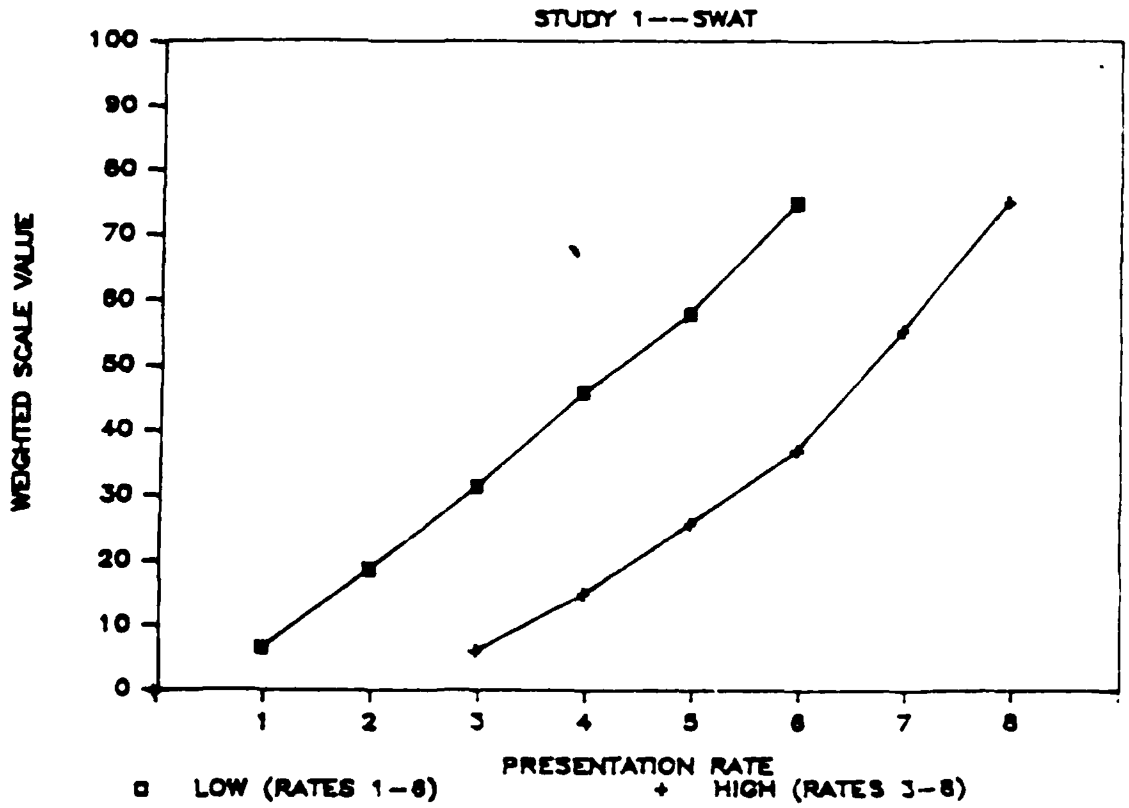


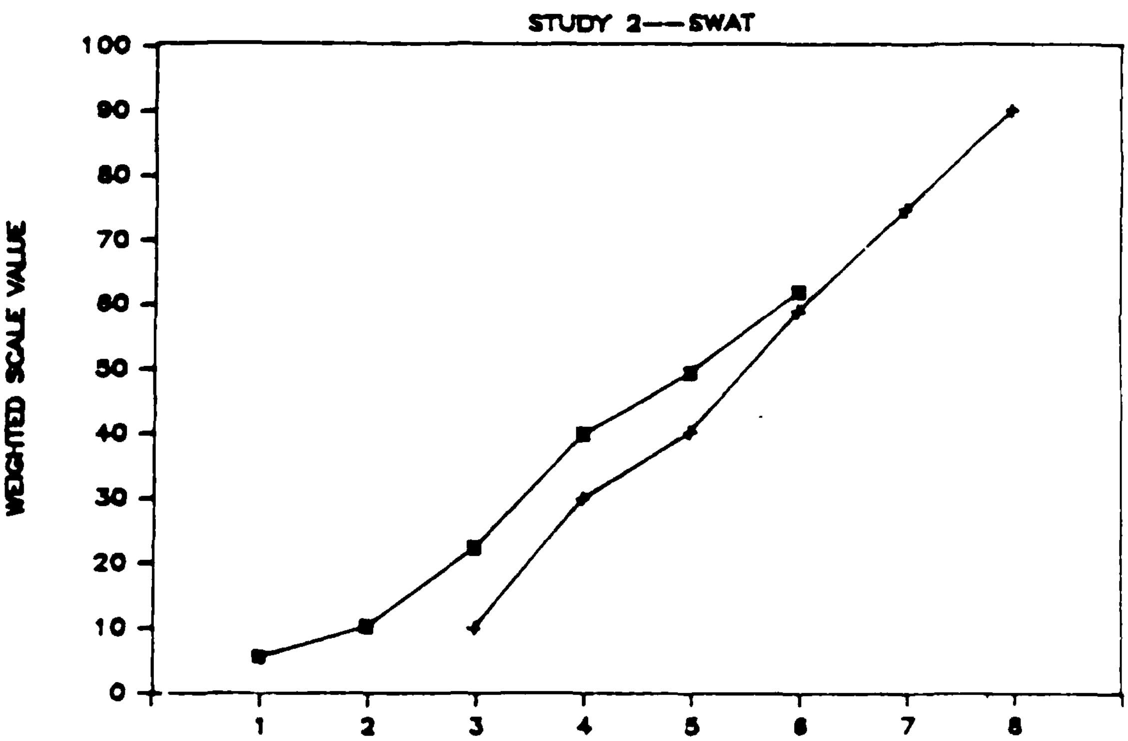
SUBJECTIVE RATINGS: MEASUREMENT ERROR
CONTEXT EFFECTS

AVERAGE WORKLOAD RATINGS OBTAINED IN DIFFERENT EXPERIMENTS (CONT):

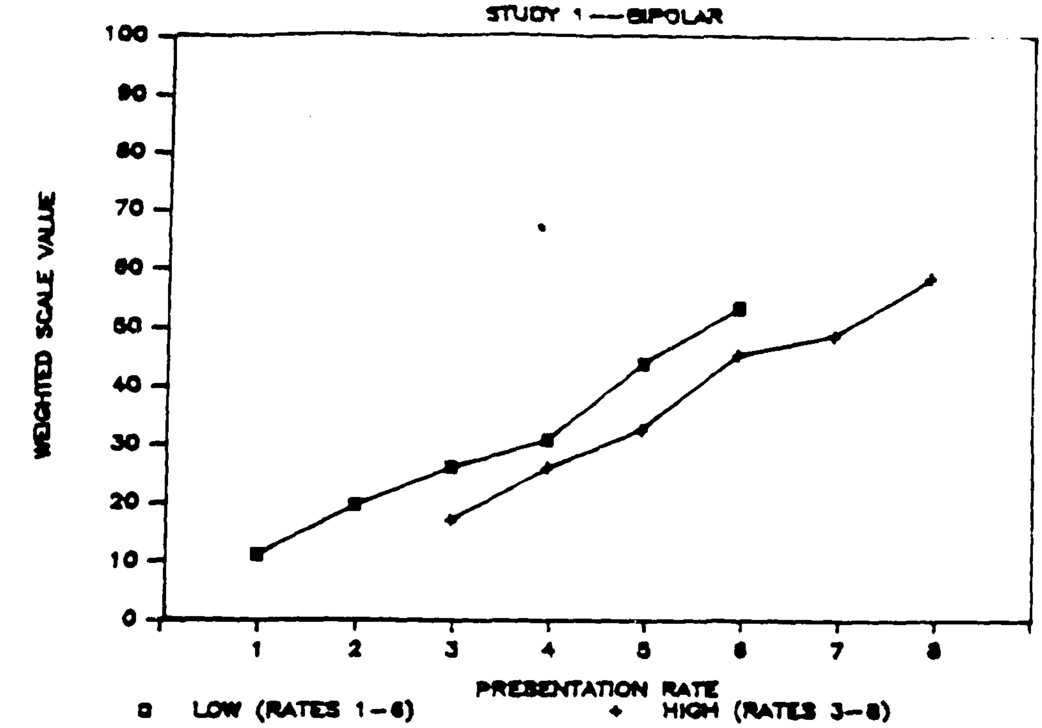


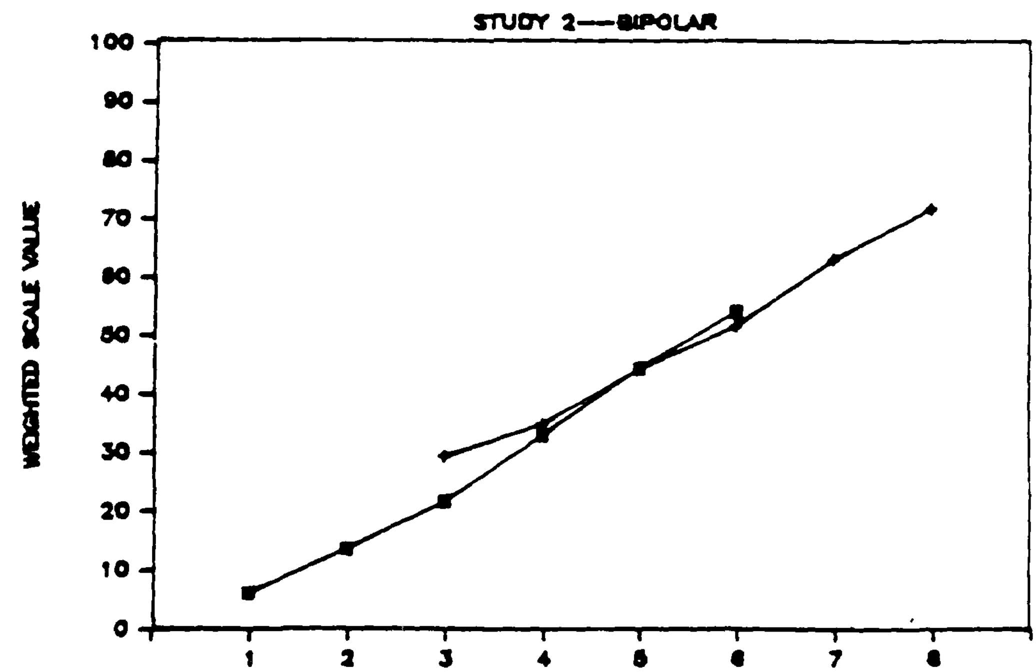
CONTEXT STUDY





CONTEXT STUDY





TOPICS DISCUSSED

1. RELIABILITY
2. SENSITIVITY
3. METHODOLOGICAL PITFALLS
   a. CONTEXT
   b. MEMORY
   c. INDIVIDUAL DIFFERENCES

RELIABILITY MEASURES

TRAITS: STABILITY OF INDIVIDUALS ACROSS DIFFERENT TASKS

STATES: STABILITY OF TASK RATINGS ACROSS DIFFERENT GROUPS OF INDIVIDUALS

VALIDITY -- WHAT IS THE MEANING OF NUMBERS?

1. WHAT ASPECTS OF THE OPERATOR / TASK INTERACTION ARE COVERED?

2. HOW DO YOU INTERPRET DIFFERENCES BETWEEN TASKS?

THE COSTS OF WORKLOAD

BEHAVIORAL MEASURES → IMPAIRMENTS TO PERFORMANCE

PHYSIOLOGICAL MEASURES → PSYCHOSOMATIC EFFECTS OF STRESS, OCCUPATIONAL DISEASES

SUBJECTIVE MEASURES → ?

SUBJECTIVE MEASURES – CONSCIOUS EXPERIENCE

1. ESTIMATES OF THE ABILITY TO COPE WITH GOALS, ACHIEVE CRITERIA

2. SENSITIVE TO THE GENERAL "WORK OF INTENTIONS"

THESE INFLUENCE PERFORMANCE ON THE VERY GENERAL LEVEL

COSTS OF SUBJECTIVE WORKLOAD

MAY EFFECT:

1. MISJUDGMENT MAY AFFECT SELECTION OF GOALS AND CRITERIA

2. MOTIVATION

3. RISK-TAKING BEHAVIOR

DEFINITIONS

"MENTAL WORKLOAD MAY BE VIEWED AS THE DIFFERENCE BETWEEN CAPACITIES OF THE INFORMATION-PROCESSING SYSTEM THAT ARE REQUIRED FOR TASK PERFORMANCE TO SATISFY PERFORMANCE EXPECTATIONS AND THE CAPACITY THAT IS AVAILABLE AT ANY GIVEN TIME." GOPHER AND DONCHIN, 1986.

"THE CONSTRUCT OF SPARE CAPACITY, DERIVED FROM MODELS OF ATTENTION, IS THE MOST IMPORTANT COMPONENT OF MENTAL WORKLOAD..."

"HOWEVER, MENTAL WORKLOAD IS MORE THAN JUST SPARE CAPACITY. ADDITIONAL ASPECTS OF MENTAL WORKLOAD INCLUDE SUBJECTIVE FEELINGS, EFFORT, INDIVIDUAL DIFFERENCES, STRATEGY AND PRACTICE." KANTOWITZ, 1986.

"WORKLOAD IS FUNDAMENTALLY DEFINED IN TERMS OF THIS RELATION BETWEEN RESOURCE SUPPLY AND TASK DEMAND." WICKENS, 1984.

"WORKLOAD ASSESSMENT TECHNIQUES ARE PRINCIPALLY DESIGNED TO MEASURE THE DEGREE OF OPERATOR PROCESSING CAPACITY WHICH IS EXPENDED IN PERFORMING A PARTICULAR TASK OR SYSTEM FUNCTION." EGGEMEIER, SHINGLEDECKER, AND CRABTREE, 1985.

"...THE DEFINITION OF WORKLOAD IN TERMS OF THE ATTENTION REQUIRED BY A TASK, OR THE ADDITIONAL CAPACITY YET REMAINING TO PERFORM ANOTHER TASK, WITH POSSIBLE REFERENCE TO THE INTENSITY OF MENTAL OR PHYSICAL EFFORT EXERTED." HART AND SHERIDAN, 1984.

"I DON'T KNOW!" "NOBODY ELSE KNOWS EITHER." KANTOWITZ, 1986.

"THERE IS NO AGREED-UPON DEFINITION OF MENTAL WORKLOAD AND NO AGREEMENT ON HOW TO MEASURE IT." MORAY, 1982.

VALIDITY

* Face Validity:

  Simulator / Aircraft

* Content Validity:

  Multiple Resources: Manual Control - Cognitive

* Construct Validity:

  Difficulty - Performance - reserve capacity

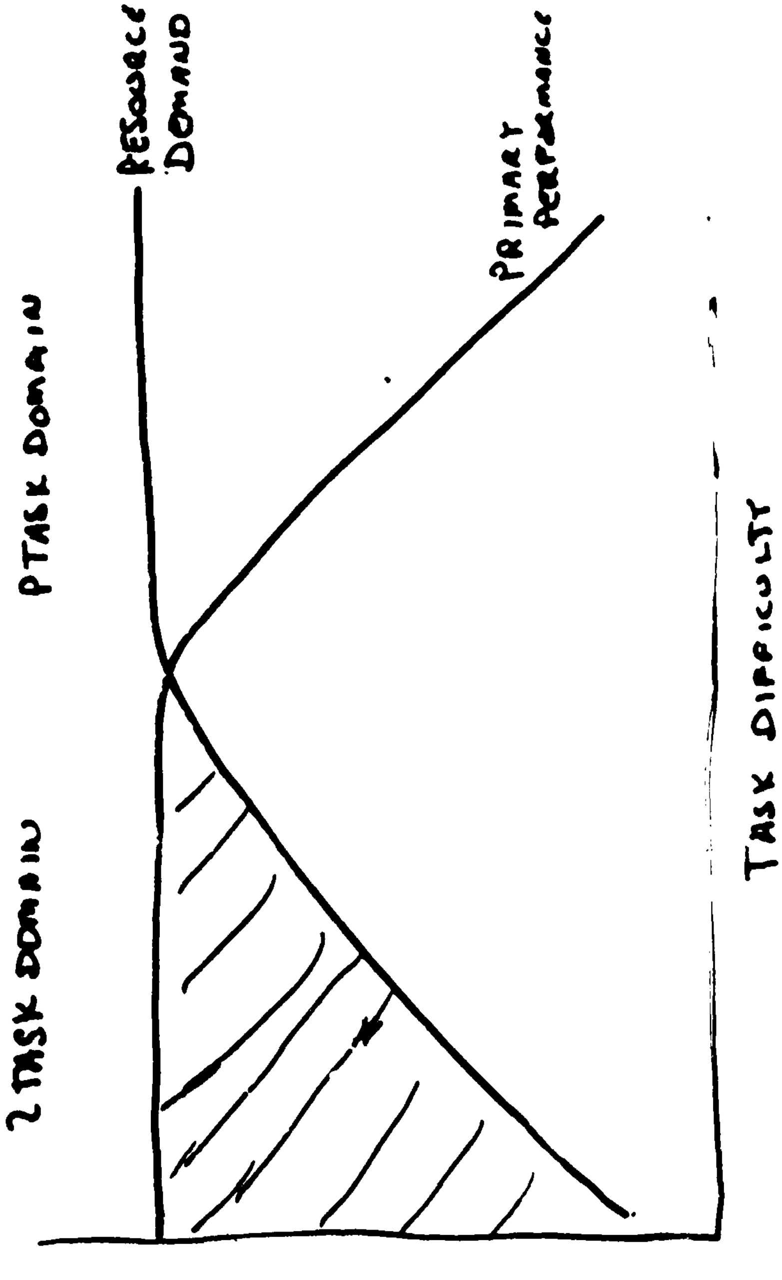
* Criterion Validity

  Prediction of Accidents

SENSITIVITY TO
MANIPULATION OF:

		PRIMARY TASK	
		Y	N
SECONDARY TASK	Y		UNDERLOAD
	N	OVERLOAD	

REJECT
2 TASK
(FOR THAT MANIPULATION)

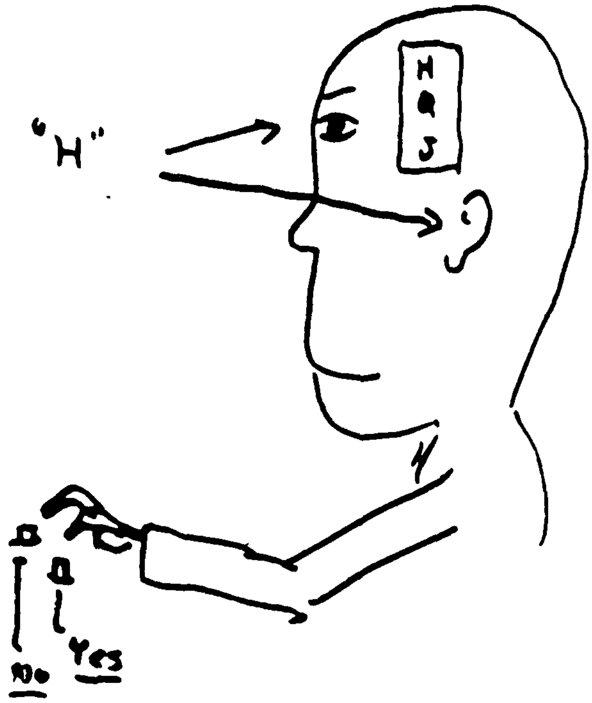

RESOURCE DEMAND
PRIMARY PERFORMANCE
P TASK DOMAIN
2 TASK DOMAIN
TASK DIFFICULTY

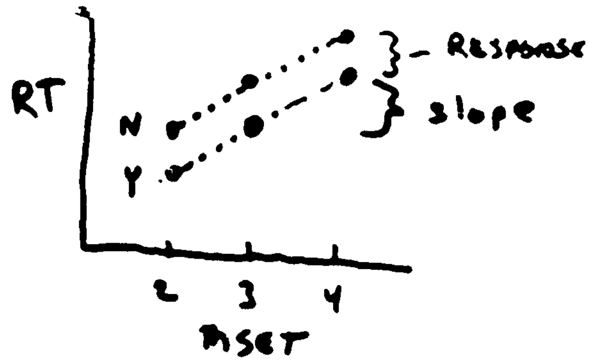
X : Sensitive
x - insensitive

	SECONDARY TASK					PRIMARY TASK		
	Sternberg	Choice RT	Cogn.	Time Est	Crit Task	Tracking Error	Control	Perf
EXPERTISE	1, 3							18
PHASE of FLIGHT	1, 6, 7	11		11*	14	16		6, 7, 11 1, 16
COMPLEXITY	5*, 4	12						12
TURBULENCE/ HANDLING	2, 8 4, 10	11	8, 13†	8	14		8	10
COGNITIVE	4				15	16, 17	8	8, 16
	9/12	2/3	1/2	1/2	3/3	1/3	1/2	5/10

* Marginal significance † Depends on modality

1. Braune & Wickens
2. Hemingway (256)
3. Crosby & Parkinson
4. Schiflett (379)
5. Spicuzza & O'Donnell
6. Dellinger } Wickens et al.
7. Hyman }
8. Wierwille et al.
9. Crawford et. al. (234)
10. Wolf (459)
11. Bortolussi et al (667)
12. Kantowitz et al. (17)
13. Sandry & Wickens
14. Jex & Clement } SCX 79
15. Clement }
16. Berg & Sheridan (358)
17. Spyker & Fort (53)
18. Stein (225)


"H"
No
Yes


RT
N
Y
Response
slope
2
3
4

Browne + W... (1984)

Table 9.

AVERAGE STATISTICS FOR EACH FLIGHT SEGMENT

Flown in GAT 2 simulator

SEGMENT	SIMULATOR CONTROL				COMMUNICATION TASK (Sternberg Task)							
	RMS	St.Dev.	Skewn.	Kurtos.	NC	St.Dev.	Skewn.	Kurtos.	RT (sec.)	St.Dev.	Skewn.	Kurtos.
I	115.65	77.54	+1.18	+1.37								
II	100.01	59.50	+2.32	+11.53	25.04	3.38	-.16	-.58	1.24	.169	-.30	-1.01
Difficult (Climb and Turn) III	193.04	140.16	+1.09	+1.23	21.52	4.56	-.45	-.72	1.35	.244	-.07	-1.02
EASY (straight and Level) IV	94.93	66.2	+2.47	+13.39	25.52	4.20	-.40	-.10	1.22	.212	+1.28	+2.93
V	106.61	65.6	+2.15	+10.61	25.37	3.85	-.33	-.94	1.22	.182	+.11	-.45
VI	187.25	122.04	+.96	+.09								
TEST-RETEST RELIABILITY (N=30)	.74				.65				.74			

STERNBERG TASK

(Wickens, Hyman, Dellinger, Taylor & Meador)

1. Difficulty of Slope Measure

2. Benefit of Y-N (Response) Difference

3. Concern for Modality
   input (Noticing)
   output (insensitivity)
   A-M is good

4. Response Switch

5. Overload Regrm (Hemingway).

6. Avoid MSET = 1

7. Avoid cm.

Source: Sardry & Wickens, 1982
F-18 Simulator Mock up

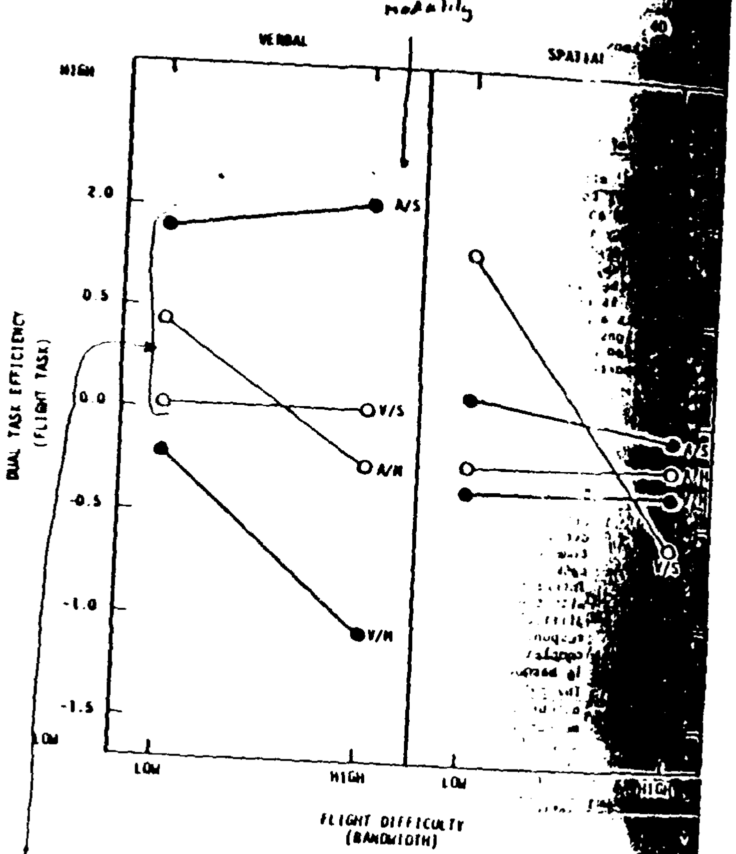


Figure 17. Flight Task Dual Task Efficiency as a Function of Difficulty

Conclusion: Voice response secondary tasks may not be sensitive to primary workload

41

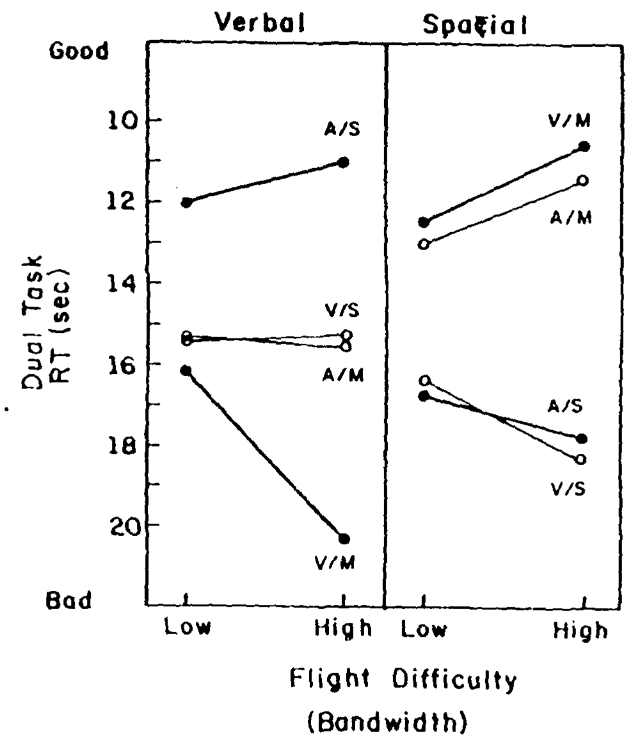
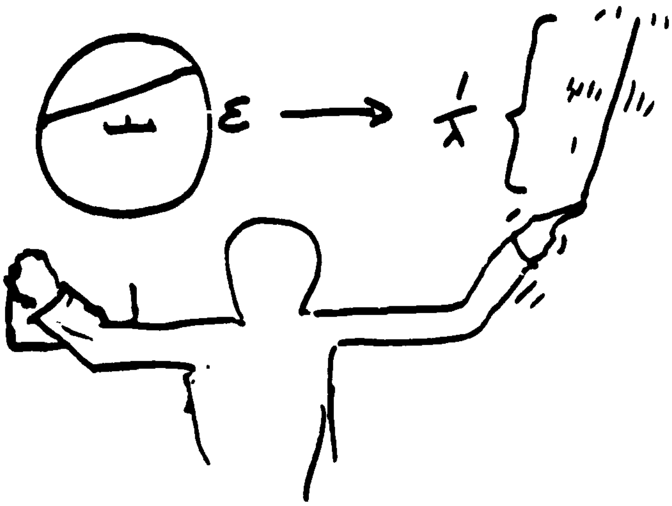


Figure 18. Dual Task Discrete Task Performance as a Function of Difficulty.

Cross-Coupled Critical Instability Task
(JEX + CLEMENT)



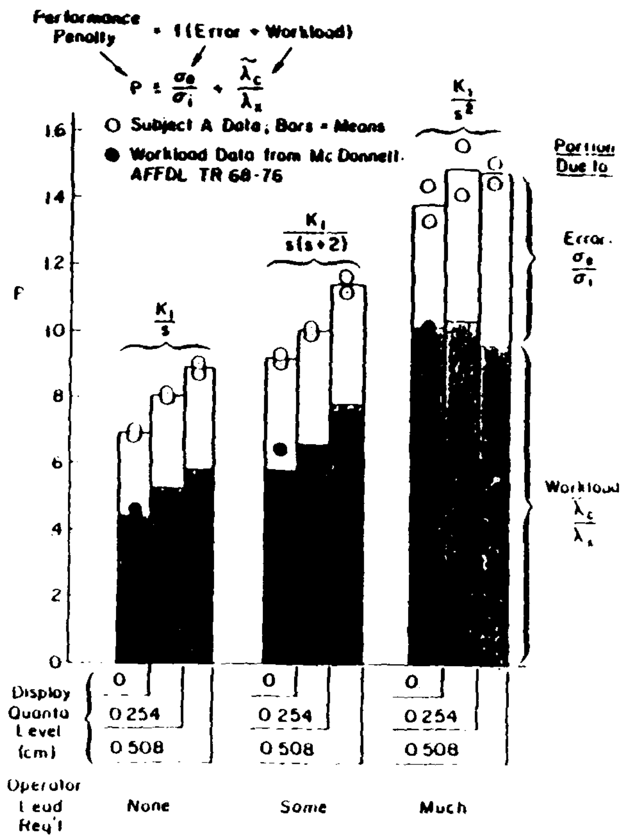


Figure 7. Typical Application of Adaptive Workload Testing (From Hess and Teichgraber, Ref. 34)

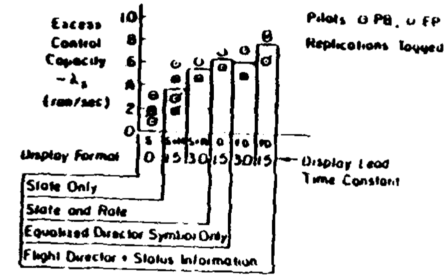


Figure 8. λ_x as a measure of excess control capacity for various VTOL integrated displays (from Clement et al., Ref. 35).

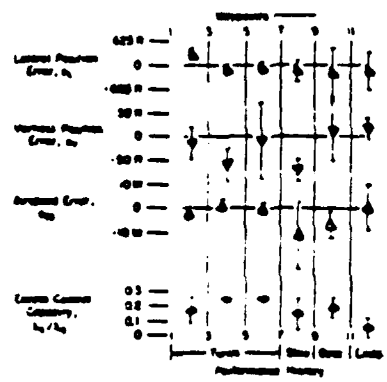
.. (Not shown) The subjective difficulty correlated with the P index.

.. Independent data on the workload index portion of P (i.e., λ_c/λ_x) is available from McDonnell's work (Ref. 33) and correlates very well with Ref. 34.

Such combined measures of error and workload can help to untangle the often anomalous trends in pure performance measures.

A third application of the CCIT in complex display research is given in Ref. 30 and 35, in which various flight director laws and integrated displays were tested with varying lead equalization time constants. The primary task was pitch and speed control during a simulated helicopter landing approach, while the cross-coupled secondary task was simulated roll control, but with an unstable element in lieu of the actual dynamics. Some results are *shown in Fig. 8. Due to task complexity the achieved values of* λ_x are low but they still distinguish among the displays in a manner consistent with concurrently sampled subjective evaluation of display workload.

Very recently, the CCIT has been employed to compare several pilots' use of a moving-map display with their use of a horizontal situation indicator (HSI) in a STOL airport approach navigation task (Ref. 36). Here the average λ_c/λ_x was evaluated over several segments of the complex approach path shown in Fig. 9 which presents an example of results by one pilot using the HSI. The cross-coupled scheme was digitally mechanized on the NASA



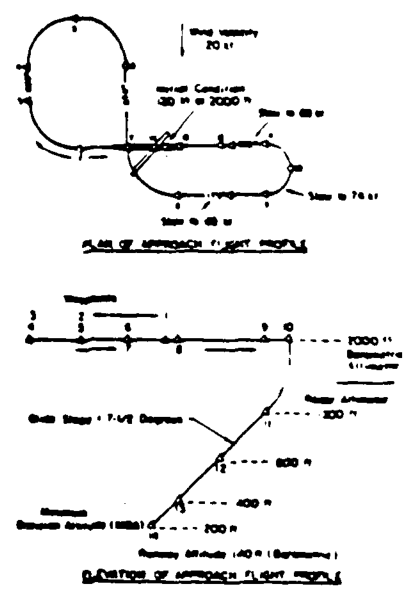


Figure 9. Sampled error performance and excess control capacity statistics with the cross-coupled task disguised as the spiral divergence of a simulated C-8M STOL aircraft. The plan elevation of the intended approach profile are also shown. The aircraft was controlled manually by a pilot using raw situation data under instrument flight rules without any flight director. Each symbol represents the time-averaged mean value and plus or minus one root-mean-square value between the numbered waypoints on the approach profile. The adaptive spiral divergence was cross-coupled to a weighted linear combination of the three error performance measures shown, and the weighted error reference was ten per cent greater than the pilot's own baseline error performance without the cross-coupled adaptive loading task. The maximum possible value of the excess control capacity measurement was limited at 0.25. This maximum value was reached between waypoints 3 and 7. (From Clement, ref. 36)

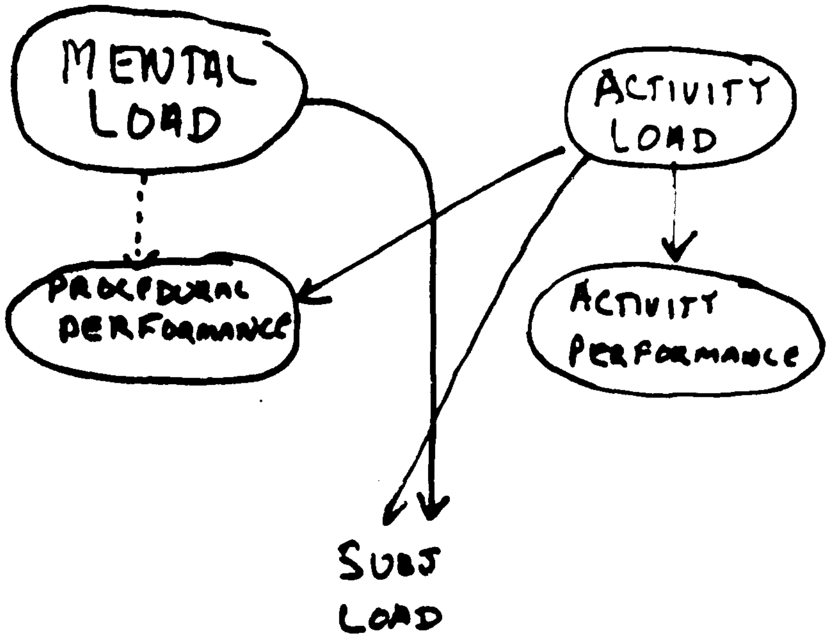
PRIMARY TASK PERFORMANCE

* DISSOCIATION with flight Phase
  (AND with Predictor information)

* INSENSITIVITY TO COGNITIVE LOAD

* ROLE OF CONTROL ACTIVITY
  (Neglected Measure?)

BERG + SHERIDAN (85)



WORKLOAD MEASUREMENT

- PERFORMANCE-BASED ASSESSMENT TECHNIQUES
  - PROVIDE MOST DIRECT MEASURES OF TWO IMPORTANT VARIABLES IN CERTIFICATION
    - LEVELS OF PILOT PERFORMANCE
    - SPARE PROCESSING CAPACITY

WORKLOAD METRIC EVALUATION CRITERIA

- VALIDITY
- RELIABILITY
- SENSITIVITY
- DIAGNOSTICITY
- INTRUSIVENESS

PRIMARY TASK MEASURES

- VALIDITY : ACCEPTABLE ON A NUMBER OF DIMENSIONS
  - CONSTRUCT
  - FACE
- RELIABILITY : NO SYSTEMATIC DATA BASE IN WORKLOAD APPLICATIONS

PRIMARY TASK MEASURES

- SENSITIVITY : RESULTS HAVE BEEN MIXED
  - DISCRIMINATE OVERLOAD FROM NONOVERLOAD
  - DISCRIMINATE WORKLOAD LEVELS IN OVERLOAD REGION
- DIAGNOSTICITY : USUAL APPLICATIONS DO NOT PROVIDE MEANS TO IDENTIFY LOCUS OF OVERLOADS
- INTRUSIVENESS : ASSUMED TO BE NONINTRUSIVE

REPRESENTATIVE APPLICATIONS OF PRIMARY TASK MEASURES IN AVIATION OR RELATED ENVIRONMENTS

- SCHULTZ, NEWELL, & WHITBECK (1970)
- HUDDLESTON & WILSON (1971)
- KRAUS & ROSCOE (1972)
- KREIFELDT, PARKIN, ROTHSCHILD, & WEMPE (1976)
- KREBS & WINGERT (1976)
- WOLFE (1978)
- NORTH, STACKHOUSE, & GRAFFUNDER (1979)

REPRESENTATIVE APPLICATIONS OF PRIMARY TASK MEASURES IN AVIATION OR RELATED ENVIRONMENTS

- CASALI & WIERWILLE (1983)
- KANTOWITZ, HART, & BORTOLUSSI (1983)
- WIERWILLE & CONNER (1983)
- CASALI & WIERWILLE (1984)
- KANTOWITZ ET AL. (1984)
- WIERWILLE, RAHIMI, & CASALI (1985)

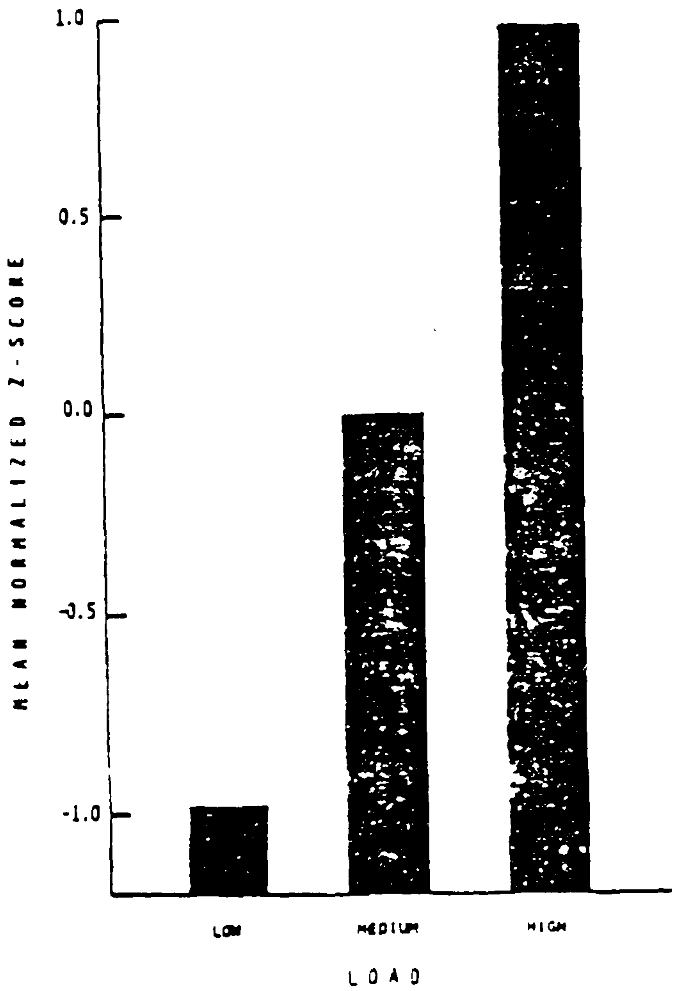
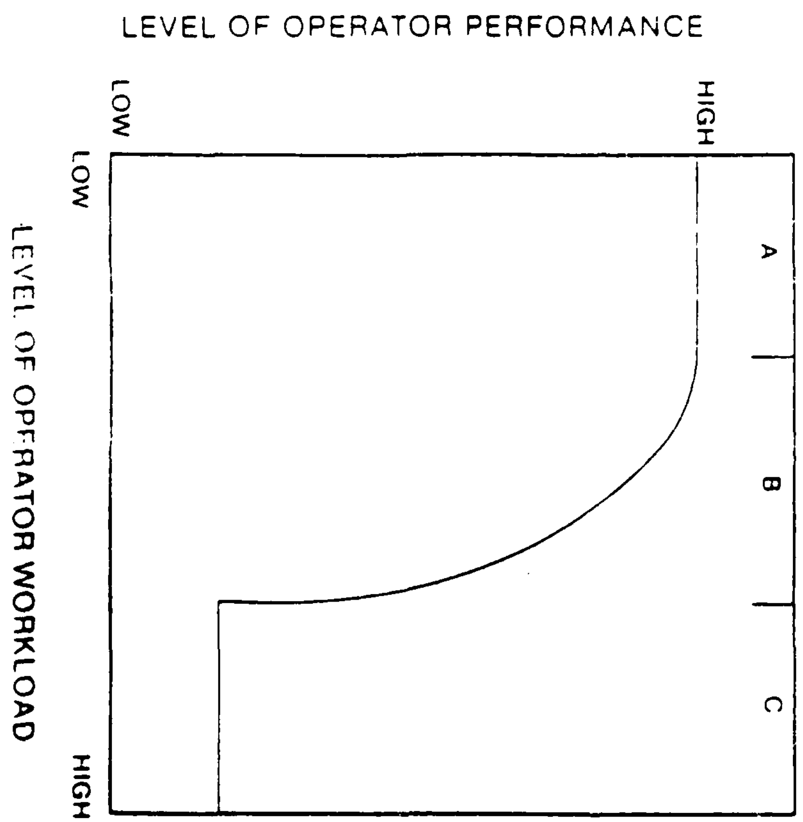


Figure 9. *Mean normalized scores as a function of load for the control movements measure.*

SOURCE: WIERWILLE & CONNER (1983)


LEVEL OF OPERATOR PERFORMANCE
LOW
HIGH
A
B
C
LOW
HIGH
LEVEL OF OPERATOR WORKLOAD

SECONDARY TASK METHODOLOGY

- VALIDITY
  - ACCEPTABLE CONSTRUCT VALIDITY
  - FACE VALIDITY COULD REPRESENT PROBLEM
- RELIABILITY
  - LITTLE SYSTEMATIC DATA IN SECONDARY-TASK APPLICATIONS
  - SOME DATA ON SINGLE-TASK VERSIONS OF SEVERAL TECHNIQUES (MEMORY SEARCH, REACTION TIME)

SECONDARY TASK METHODOLOGY

- SENSITIVITY : CAPABLE OF DETECTING CAPACITY EXPENDITURE VARIATIONS IN NONOVERLOAD CONDITIONS

- DIAGNOSTICITY : CAN PROVIDE MEANS TO SPECIFY LOCUS OF POTENTIAL OVERLOADS

- INTRUSIVENESS :

  - HAS REPRESENTED PROBLEM IN LABORATORY

  - PROBLEM MAY NOT BE AS SEVERE IN SIMULATION/OPERATIONAL ENVIRONMENTS

REPRESENTATIVE APPLICATIONS OF SECONDARY TASK METHODOLOGY IN AVIATION OR RELATED ENVIRONMENTS

- STERNBERG MEMORY SEARCH
  - O'DONNELL (1976)
  - CRAWFORD, PEARSON, & HOFFMAN (1978)
  - WOLFE (1978)
  - WICKENS & DERRICK (1981)
  - SCHIFFLET, LINTON, & SPICUZZA (1982)
  - WIERWILLE & CONNOR (1983)
  - HEMINGWAY (1984)
  - WICKENS, HYMAN, DELLINGER, TAYLOR, & MEADOR (1985)

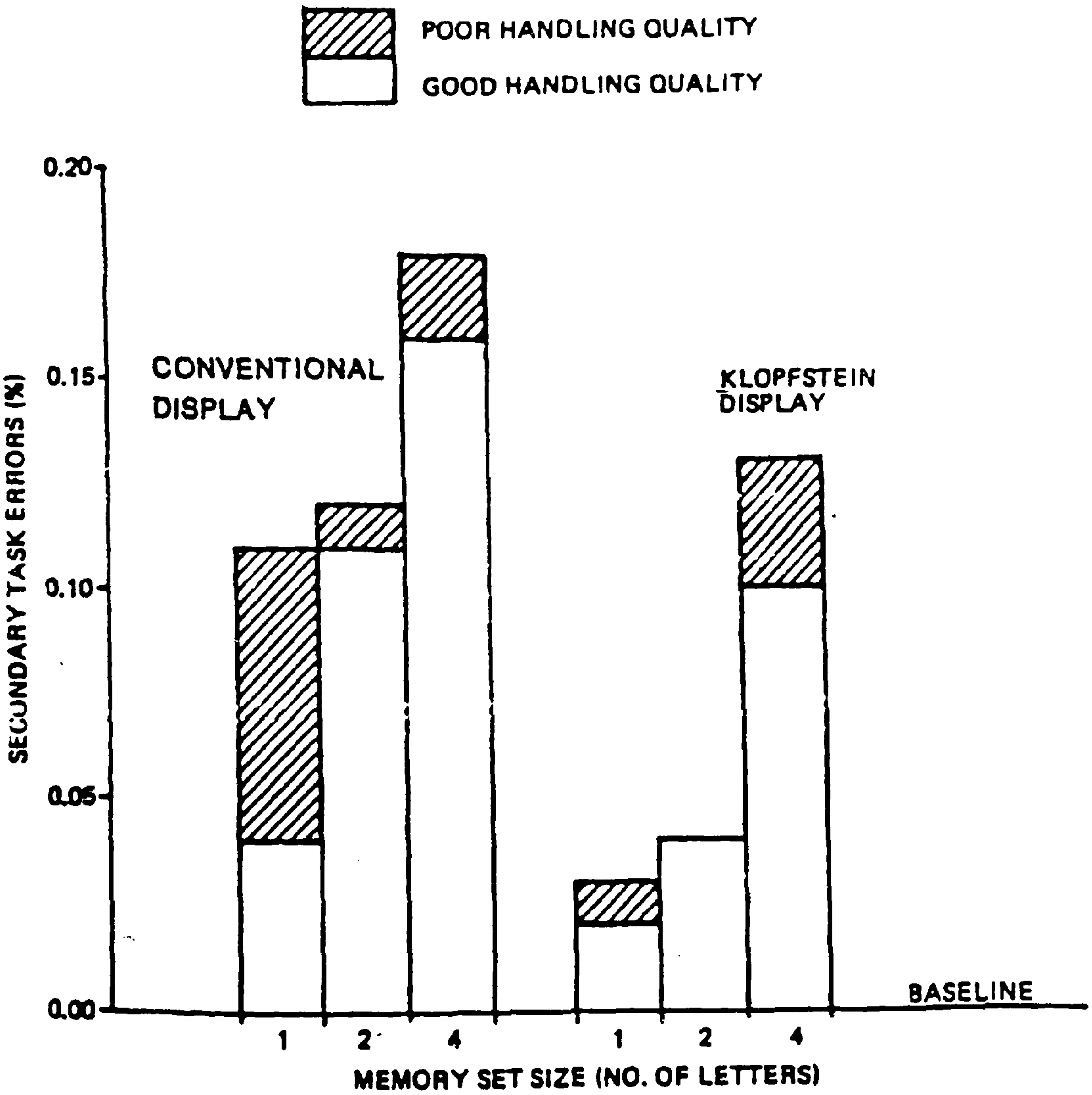


Figure 6 **Mean Percent Secondary Task Error for Memory Set Size (Number of Letters) by Display Format and Handling Quality**

SOURCE: SCHIFFLET, LINTON, & SPICUZZA (1982)

REPRESENTATIVE APPLICATIONS OF SECONDARY TASK METHODOLOGY IN AVIATION OR RELATED ENVIRONMENTS

- CHOICE REACTION TIME
  - KANTOWITZ, HART, & BORTOLUSSI (1983)
  - KANTOWITZ ET AL. (1984)
- MENTAL MATHEMATICS
  - HUDDLESTON & WILSON (1971)
  - GREEN & FLUX (1976)
  - WIERWILLE & CONNOR (1983)

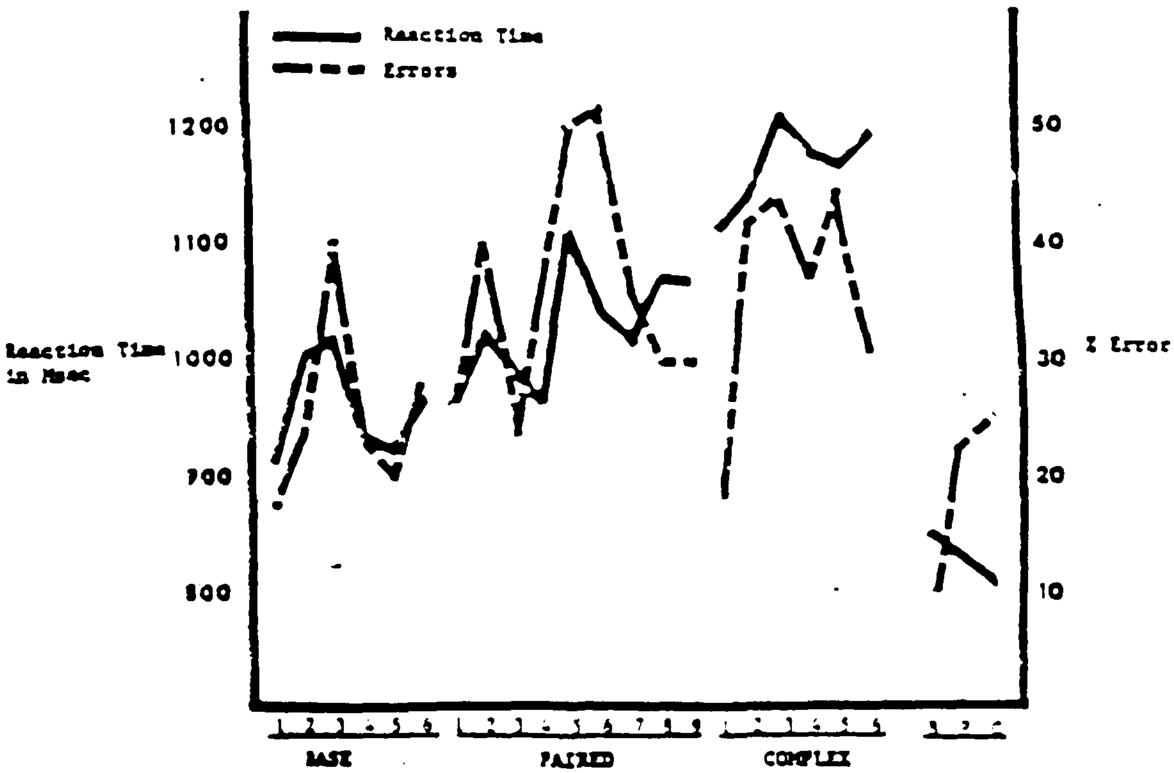


Figure 5. Secondary task performance (Reaction time/Errors) as a function of level of primary task.

KANTOWITZ, HART, BORTOLUSSI, SHIVELY, & KANTOWITZ (1984)

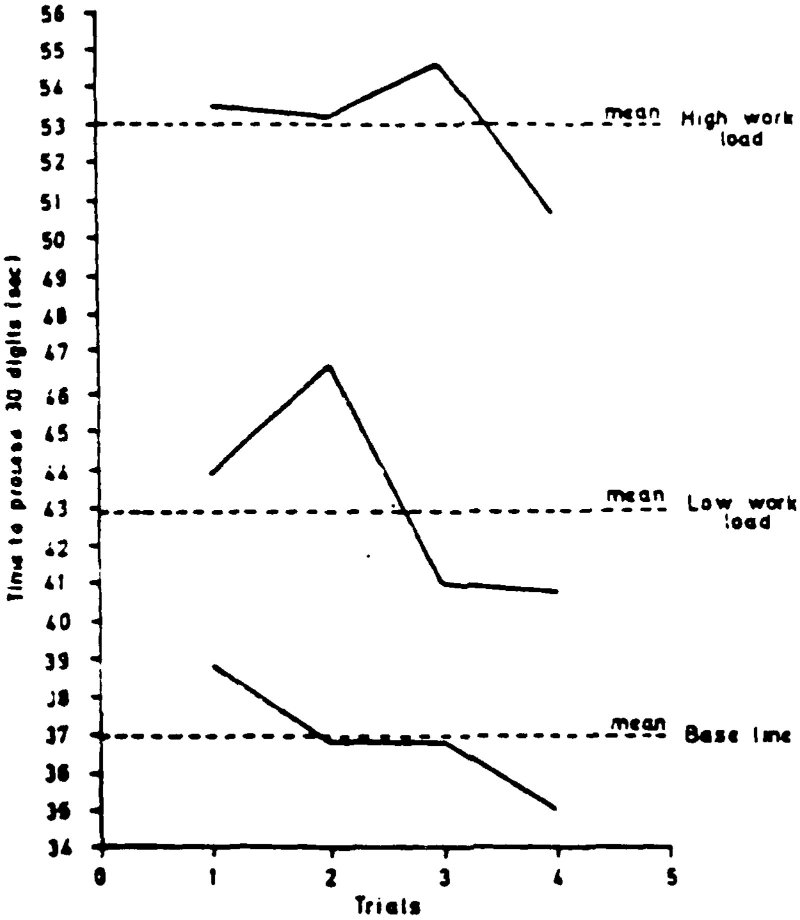


Fig3 Effect of Cockpit Workload on a Auditory Task

SOURCE: GREEN & FLUX (1976)

REPRESENTATIVE APPLICATIONS OF SECONDARY TASK METHODOLOGY IN AVIATION OR RELATED ENVIRONMENTS

- CRITICAL TRACKING TASK
  - JEX & CLEMENT (1979)
  - BURKE, GILSON, & JAGACINSKI (1980)
- TIME ESTIMATION
  - HART (1978)
  - GUNNING (1978)
  - WIERWILLE & CONNOR (1983)
  - CASALI & WIERWILLE (1983)

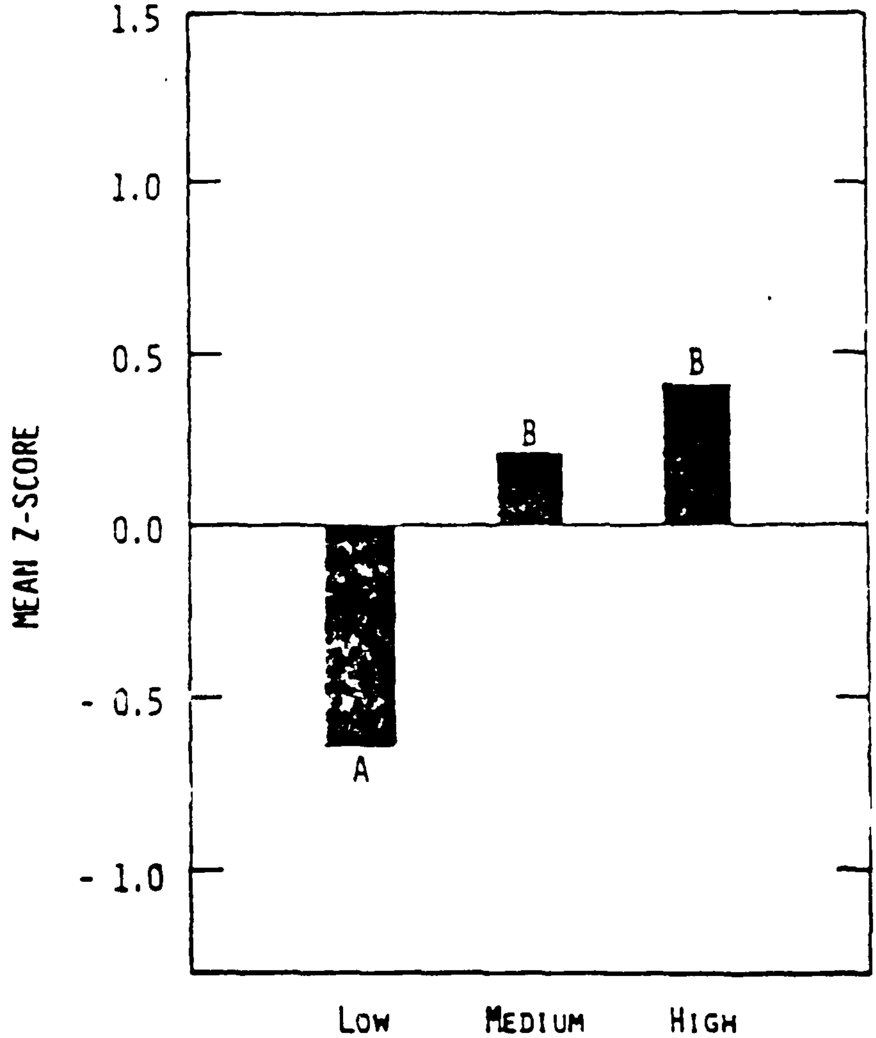


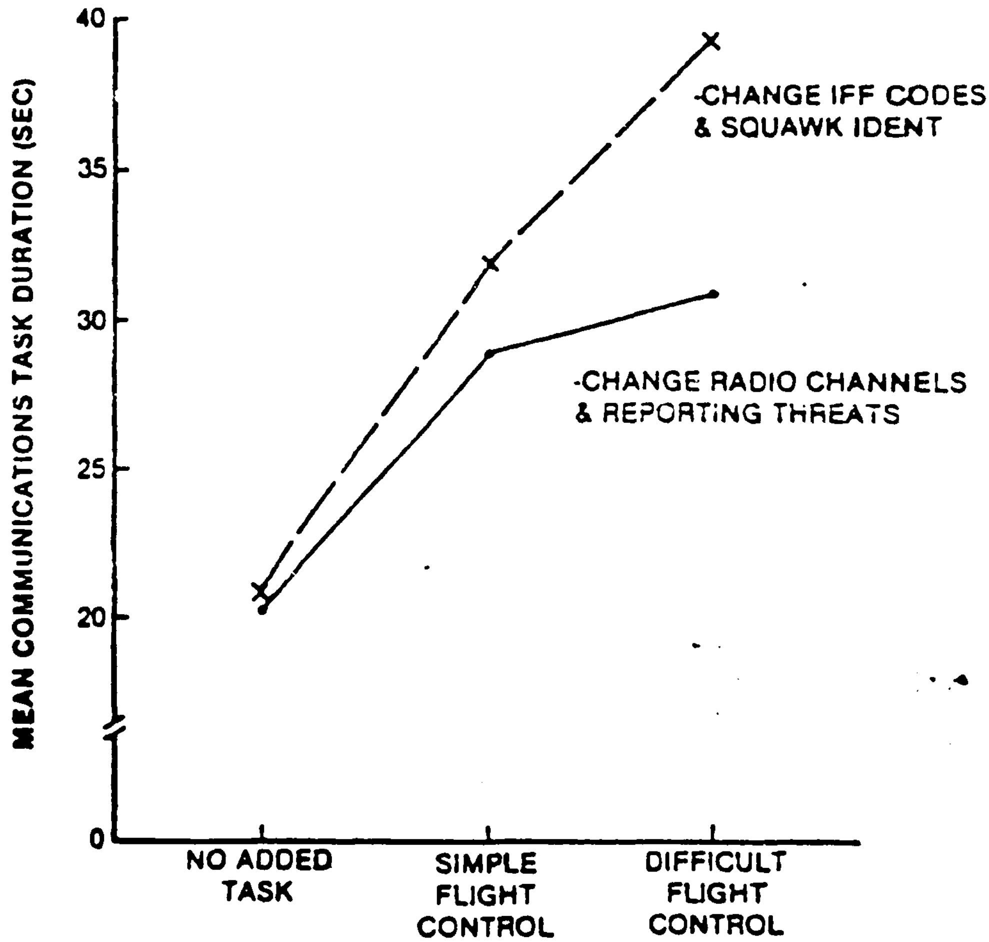
Figure 3. *Effect of load on mean standardized scores for the time estimation standard deviation technique. (Means with different letters are significantly different,* $p < 0.05$*).*

SOURCE: CASALI & WIERWILLE (1983)

REPRESENTATIVE APPLICATIONS OF SECONDARY TASK METHODOLOGY IN AVIATION OR RELATED ENVIRONMENTS

- EMBEDDED RADIO COMMUNICATIONS
  - SHINGLEDECKER ET AL. (1980)
  - SHINGLEDECKER & CRABTREE (1982)
  - SILVERSTEIN, GOMER, CRABTREE, & ACTON (1984)

COMMUNICATION TASK PERFORMANCE AS A MEASURE OF COCKPIT WORKLOAD



SOURCE: SHINGLEDECKER & CRABTREE (1982)

CONCLUSIONS

- PRIMARY TASK MEASUREMENT ESSENTIAL IN ALL APPLICATIONS
- SECONDARY TASK MEASURES PROVIDE IMPORTANT COMPLEMENTARY INFORMATION THROUGH:
  - POTENTIAL GREATER SENSITIVITY TO VARIATIONS IN CAPACITY EXPENDITURE
  - CAPABILITY TO DIAGNOSE THE LOCUS OF POTENTIAL OVERLOADS
- SECONDARY TASK BATTERY REQUIRED TO PROVIDE DIAGNOSTIC CAPABILITY

- Primary Task Measures
- Secondary Task Measures
  - Sensitivity
  - Diagnosticity
  - Intrusiveness
    - User Acceptance
- Underload
- Battery

Issues

1 Performance Workload Measures

- Primary – Necessary but not sufficient – bottom line is can you do the job accomplish the task.
  - Sensitivity an issue especially at the low end and in the middle. At the high end, if the operator can no longer perform it may be indicative of a workload problem depending on the other circumstances.
  - But what were the costs to this perf?
- Secondary – Diagnostic some appear to offer promise at discerning in a valid & reliable fashion, costs of performance in terms of resources etc

Sternberg – flexible
Critical tracking task

• Underload – most of the literature we address are not focused on this – may be different sides of the coin. For example looking at the analogy of attention & anomalies of attention – High workload frequently results in over focusing of attention (tunnel vision) whereas low workload may be more of boredom and lack of the proper attention — distraction

• Fidelity & Pilot Acceptance –

• A criticism of 2nd tier procedures

• ~~Class~~ Implementation problem of Intrusion

• embedded tasks seen as promising

• Crew vs Pilot Workload

① Physiological Measures of Workload

Ⓘ "WORKLOAD"

(Task Demands - Operator Responses)

"Costs"

Physical	Affective	Mental
* general metabolic costs	* emotions	Taxonomies:
* specific costs	* acute vs chronic	(a) Wierwille & Casali:
* static vs. dynamic		* psychomotor
		* perceptual
		* mediational
		* communication

(b) Wickens, Navon & Gopher, Friedman & Polson

Resources:

* stages
* codes
* modalities

(c) Sanders

Resource - Stage model

3 II Why Physiological Measures?

Considerations:

* Cost
    training time & money
* Equipment
    transducers, amplifiers, computers
* Signal recognition - Signal Extraction - Signal conditioning
* Fact vs Artifact
    ERPs — EOG, EMG, 60/50 Hz
    Frequency, Spatial & Temporal Filters

So . . . .

Physiological measures should complement / augment other measures

(3)

A few benefits:

* unobtrusive

* multidimensional

  * Cardiac function - EKG, blood pressure, blood volume, O_2

  * CNS function - EEG, ERPs

* Continuous

* Real-Time

* Range of Temporal Resolution

BER's HR GSR — Glucose metabolism

msec — mins

Also address EEG & ERP

4 III Categories of Physiological Measures

A **W**ORKLOAD

heart rate, sinus arrhythmia, blood pressure, SCR, respiration, EMG, eyeblink

B **w**orkload

ERPs, EEG, pupil diameter, body fluids

DIAGNOSTICITY

Lo ———————————— Hi

Components of Workload important for Aircraft Certification

"APPLICABILITY"

(2) (IV) Specific 'Physiological' Measures: Problems & Promises

Cautionary Note: Physiological measures should not be collected in isolation. (Workload anchors / validation criteria are necessary)

add criteria also necessary for oth. measures – Multidimensional Concept

(A) Heart Rate & Heart Rate Variability

Why?

① Large data base suggesting association between heart rate & variability & affective / physical / mental workload

Construct Validity

Aviation Context:

* Alan Roscoe – flight segments
* Lindqvist et al – " "
* Hart et al – pilot / co-pilot

(6) but . . . .

"not all roses"

Failures to find association:

* Wierwille et al
* Mobbs et al

Explanations:

* Variety of calculation methods used for heart rate / heart rate variability
* Different components of "signal" (P QRS T) may be sensitive to different aspects of workload

(a) Kalsbeek et al (1960's)

* heart rate variability decreases with little change in heart rate (perceptual-motor tasks)

(b) Mulder . . . .

* .02 - .06 Hz → vasomotor activity involved in body temp. regulation
* .15 - .50 → reflects respiratory activity on cardiac signal

To the extent any physio measure is practical.

② Multi dimensional measure

"varieties of workload"

③ Easy to Record & Analyze

"Practical"?

Time / Frequency Domains

Artifacts:

(a) low frequency - convective characteristics of skin

filter

(b) high frequency - movement / muscle activity

* high freq tougher
- in submitted ok.

④ Reliability

* Consistency across experiments
alternate forms / investigators

* Test - Retest

Fahrenberg et al (1986)

30 min .81
3 weeks .64
1 year .47

→ over variety of conditions & 58 S's

9) (B) Event Related Brain Components

Why?

(1) Multidimensional Measure

* Between Components
  - Exogenous - endogenous
  - integrity of CNS - workload
* Within Component

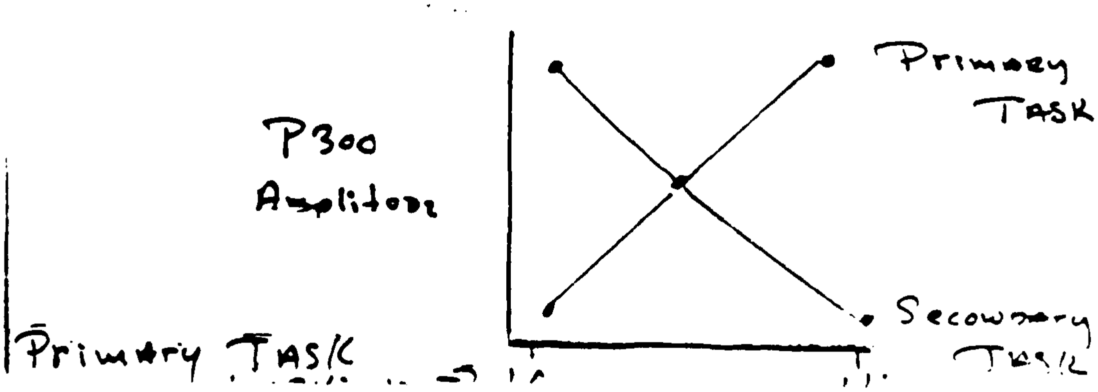
  latency, amplitude, scalp distribution

also EEG is "free"

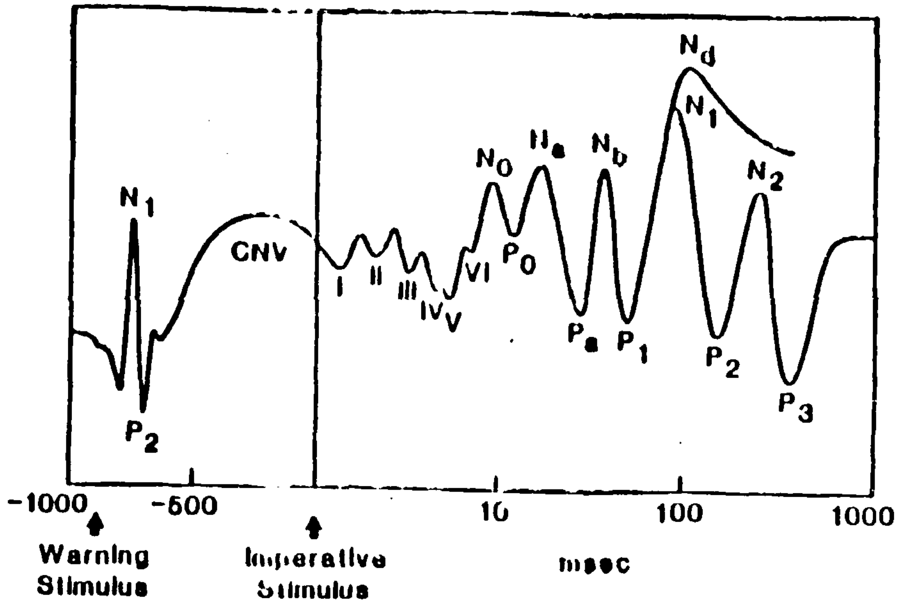
"Convergent Validity"

(2) Large data base suggesting association between perceptual/cognitive demands & characteristics of ERP (P300)

* Strong tie with models of Resource Allocation & Mental Chronometry.



Auditory Event-Related Potentials



10

Validity in Aviation Contexts

* Wierwille & Connor (1983)
    * Part Task Simulation
        maintain constant airspeed & perform threat detection task
        * Secondary task technique

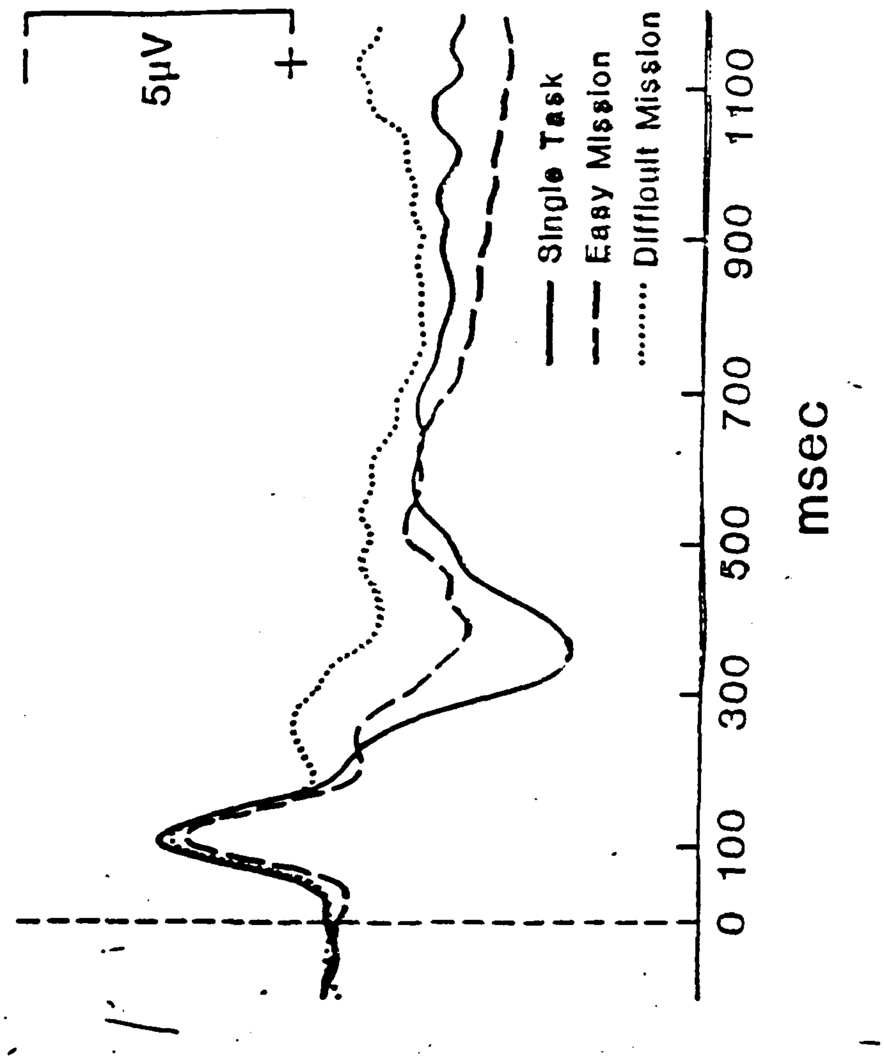
* Lindholm et al (1984)

Carrier Landing (pilot trainees) & Surface to Air missile avoidance

* # of tasks
* Threat level
* Secondary/probe technique
* Secondary/primary technique
* N200 & P300

* Kramer et al (1987)
    * Full task IFR simulation
        * Between & within mission analysis
        * Student & Instructor Pilots
        * Secondary Task Technique.

WORKLOAD
~~Pitch~~ "Turbulence": Pitch Roll
(Difficulty # of TASKS, Caution, Caution & Tone, Tone.)
# of Pilot Manuals


5µV
Single Task
Easy Mission
Difficult Mission
0 100 300 500 700 900 1100
msec

(11)

(3) Recording & Analysis

* Potential Artifacts
  * EOG - filters (EMCP)
  * EMG - influences some components more than others
  * Line noise - ground loops
    notch filters
  * S/N Ratio

Average

Single Trial Conditioning.
"Mental Prosthesis

Analysis

* Visual inspection — Univariate Measures — Multivariate Techniques

Paradigms:

Used for Discrete event . . .

Primary, Secondary, Probe

Secondary Task also useful - "transportable."

B

4 Reliability

Fabiani et al (1987) - "P300"

* split-half
* test-retest
* Comparison of = # of Data Analytic studies
* comparison of different features of component

	Amplitude	Latency
Split Half	.92	.83
Test Retest	.83	.63

(13) C BLINKS

(1) Multicomponent Measure

- Blink latency, rate
- Blink waveform
  - closing
  - closed
  - reopening

(2) Large database suggesting association between blink parameters & attentional focusing, transition points in information processing, level of activation

- sensitive to General Cognitive Factors

Mostly laboratory tasks
- Stern et al
- Sherman
- Holland & Tarrow

Keep operational Validation

(14)

③ Recording & Analysis

* Variety of Recording techniques
  * ECG is free
    (complements ERPs)
      noise – signal
      signal – noise
* Similar electrical artifact
  Problems to ERPs
  but better S/N.

* Analysis relatively standard

* Can't count on blinks
  every trial (e.g. blink rate)

  Hi ——————— Lower
  ERPs   Blink Rate
  "Temporal Resolution

④ Reliability
  * Separate Focus

(15)

SUMMARY

* Complement other measures
    * general system activation
    * diagnostic resource costs
        & chronometric information
    * automatic vs. controlled
* Other Physio measures interesting but .....
    e.g. Pupil Diameter
* unobtrusive
* continuous
    time course
* Standardized measures needed
    & recording procedures
* Provide information on the Physical integrity of the operator

PHYSIOLOGICAL MEASURES

- NON-INTRUSIVE
- CONTINUOUS
- IMMEDIATE

CANDIDATE MEASURES

- CARDIAC ACTIVITY
- BLINK ACTIVITY
- EVOKED POTENTIALS
- PUPIL DIAMETER
- RESPIRATION
- EEG ANALYSIS
- MUSCLE ACTIVITY
- BLOOD ANALYSIS
- URINE ANALYSIS

NEUROPSYCHOLOGICAL WORKLOAD TEST BATTERY (NWTB)

CENTRAL MEASURES

- TRANSIENT EVOKED RESPONSE
  - ODD-BALL
  - MEMORY SCAN
  - MEMORY UPDATE
  - SELECTIVE ATTENTION
  - MONITORING
  - BRAIN STEM
  - TRACKING
  - FLASH
- STEADY STATE EVOKED RESPONSE
  - CHECKER BOARD
  - SINE WAVE GRATING
  - UNPATTERNED FIELD

PERIPHERAL MEASURES

- EYE BLINK
- HEART RATE
- MUSCLE

PERFORMANCE MEASURES

- REACTION TIME
- ERROR SCORES

HE 85 5 32

CARDIAC ACTIVITY

- MEASURES
  - RATE
  - VARIABILITY
  - T-WAVE AMPLITUDE
- RELEVANT DATA
  - COMMERCIAL
  - MILITARY
- USES
  - STRESS
  - MENTAL ACTIVITY
- EASY TO IMPLEMENT
- GENERALLY ACCEPTABLE

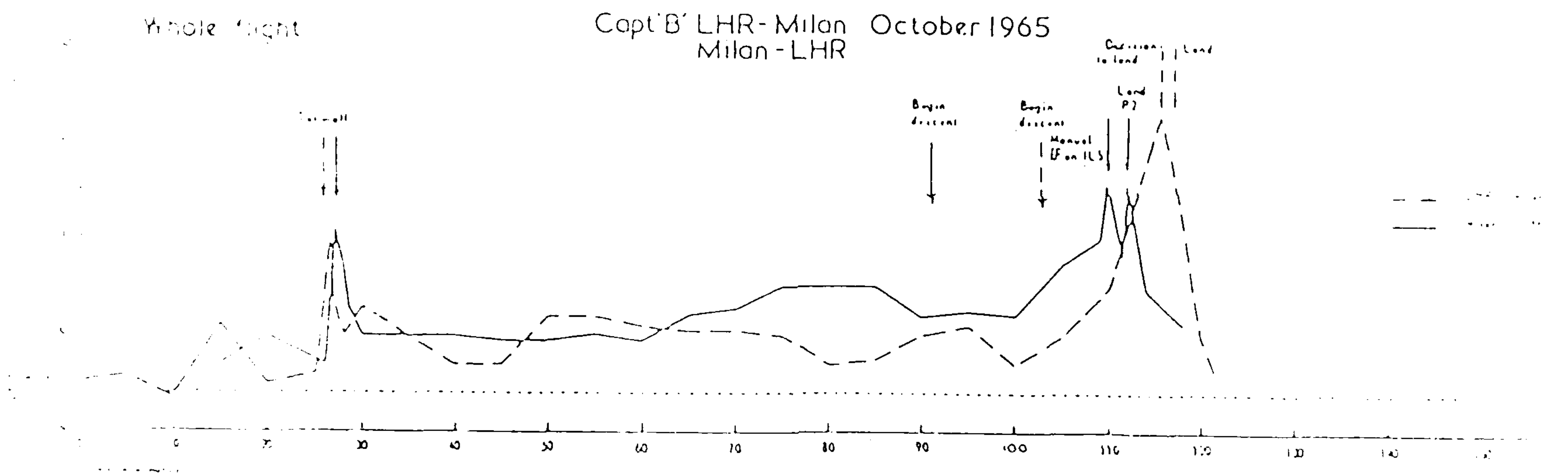
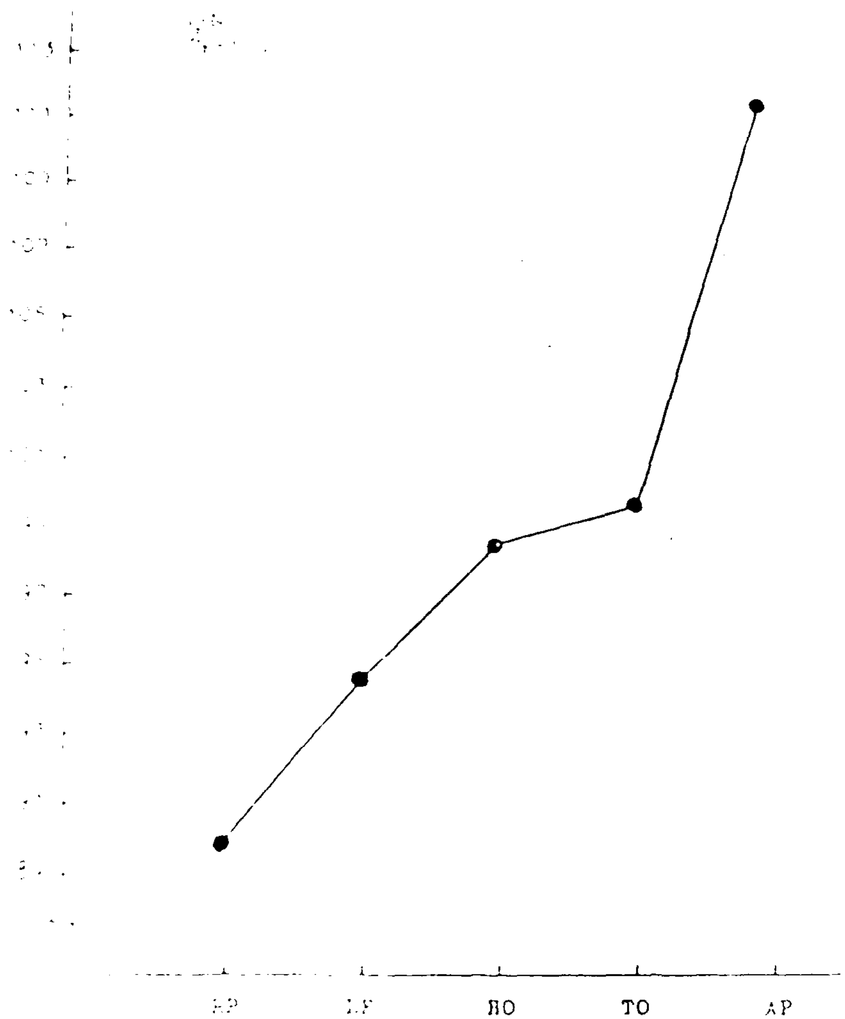


Fig. 2. Heart rate record during complete flight.

Smith, 1967



... per min of two pilots, making nine ... days in a DC-7 simulator, during ... of the ROT pattern. The difficulty of the ... from left to right.

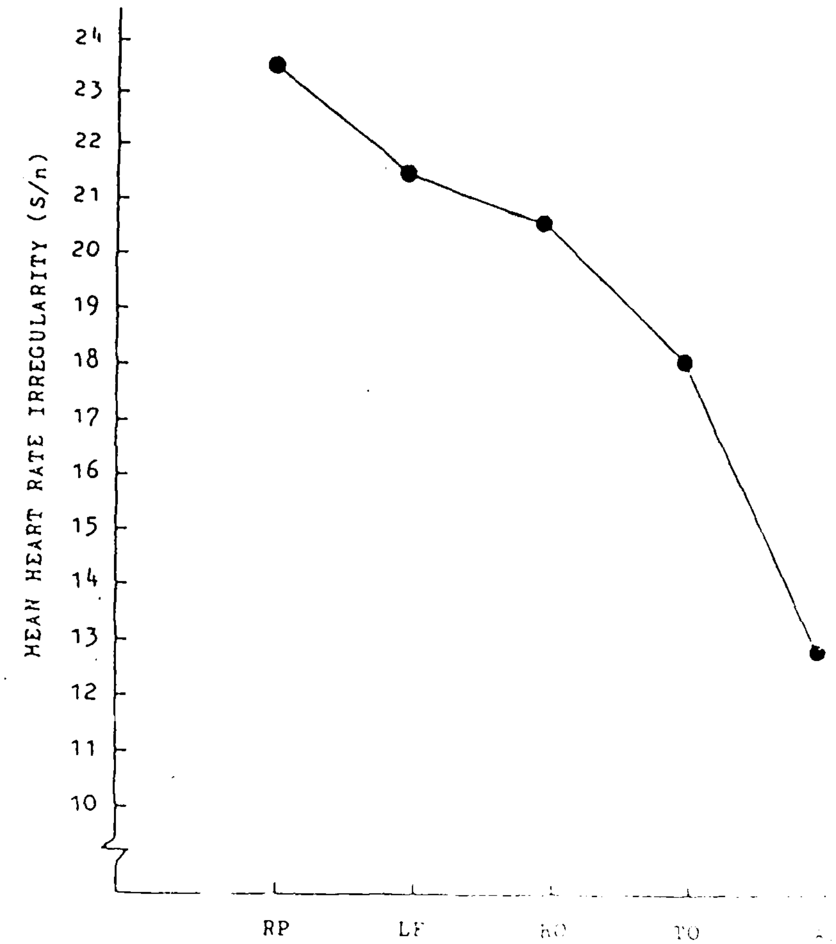


Fig. 4. Mean *heart rate irregularity* per min of ... making nine flights each on three different days in a ... simulator, during different phases of the ROT pattern ... difficulty of the phases is supposed to increase from ... right.

Opmeer, 1973

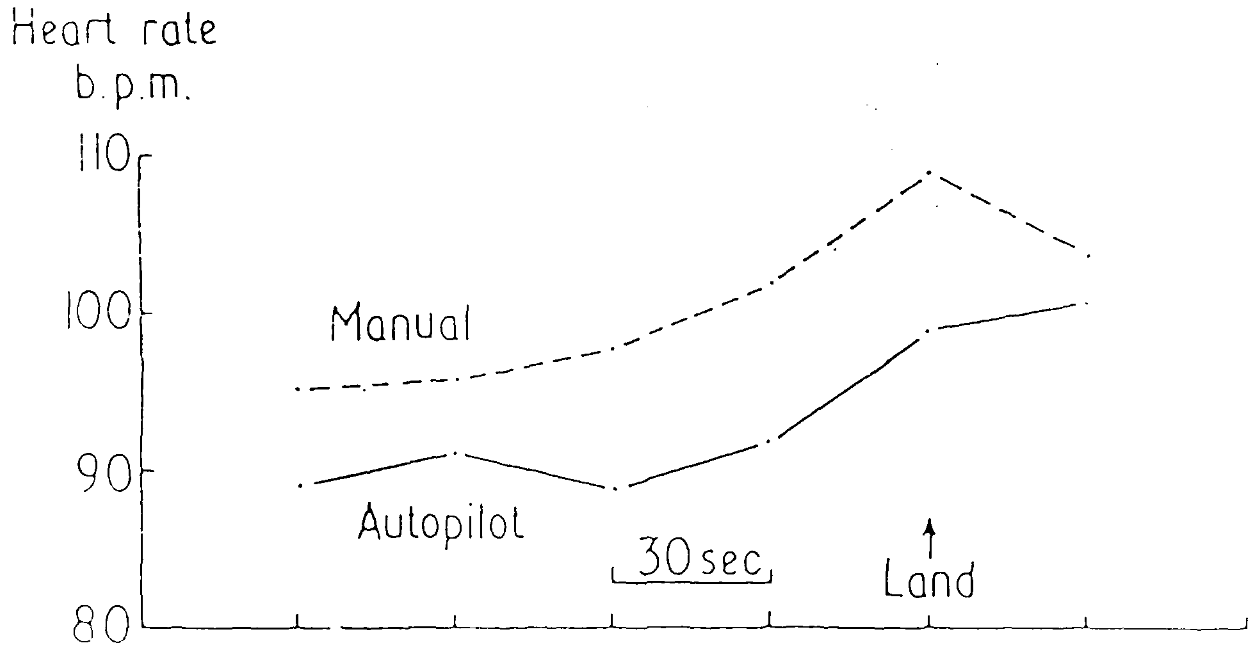
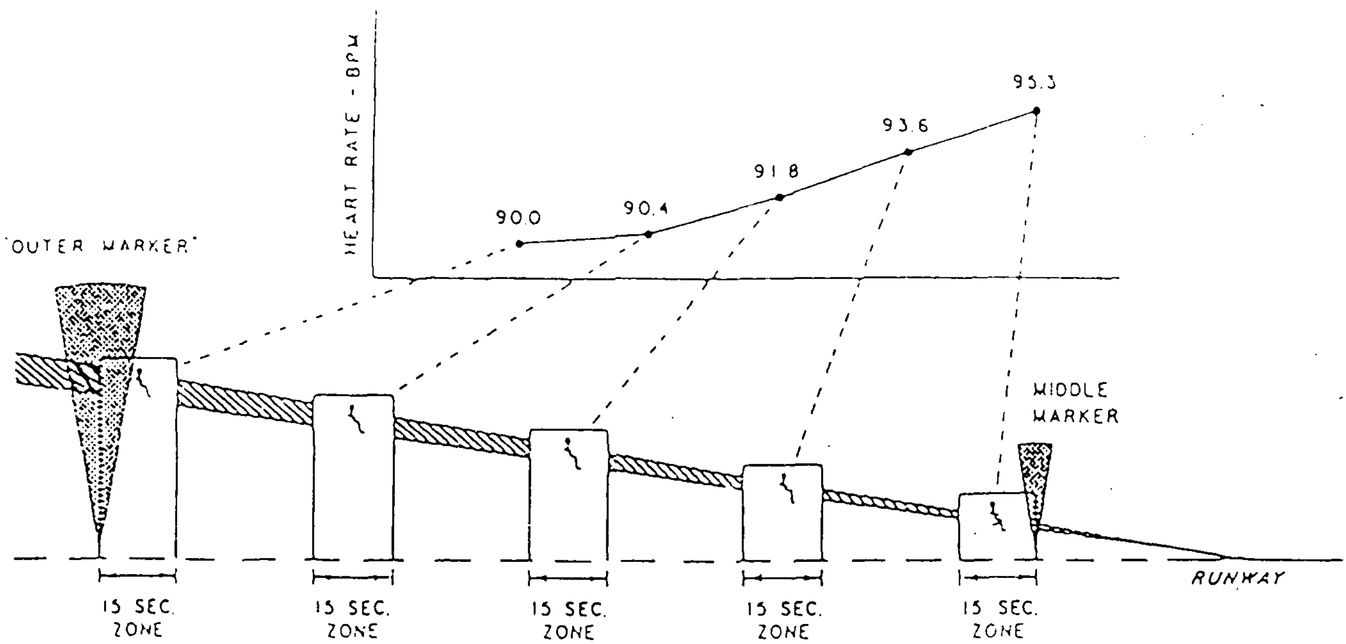
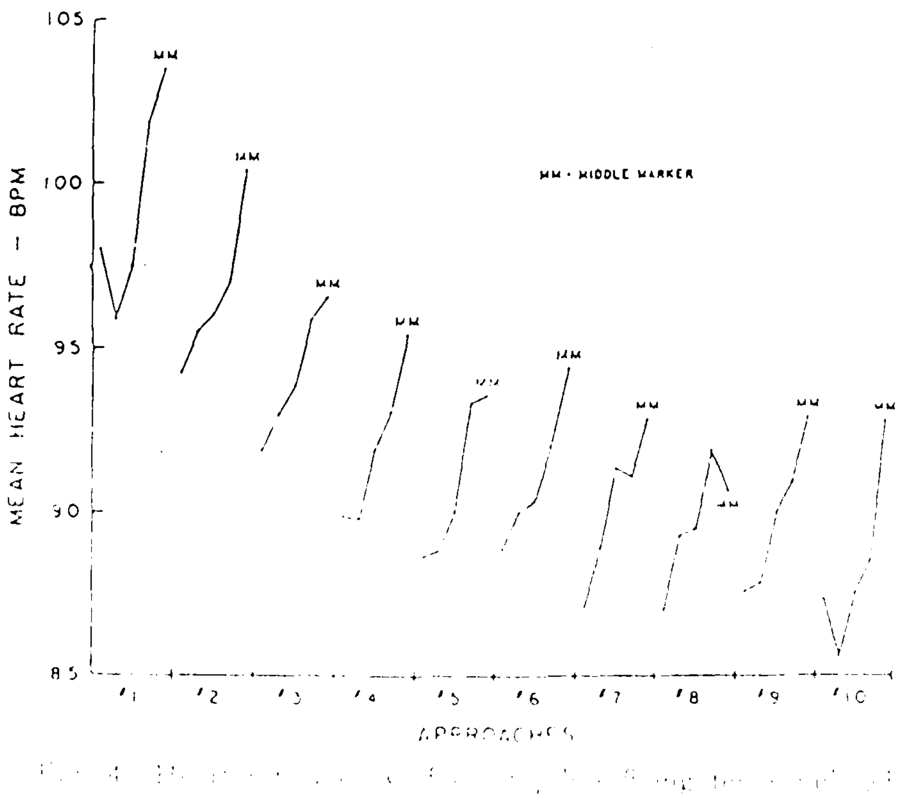


Fig. 2 Mean heart rate values (30-s epochs) for three manual and six autopilot approaches and landings during the same sortie in Comet 3B. (Pilot C)

Roscoe, 1976

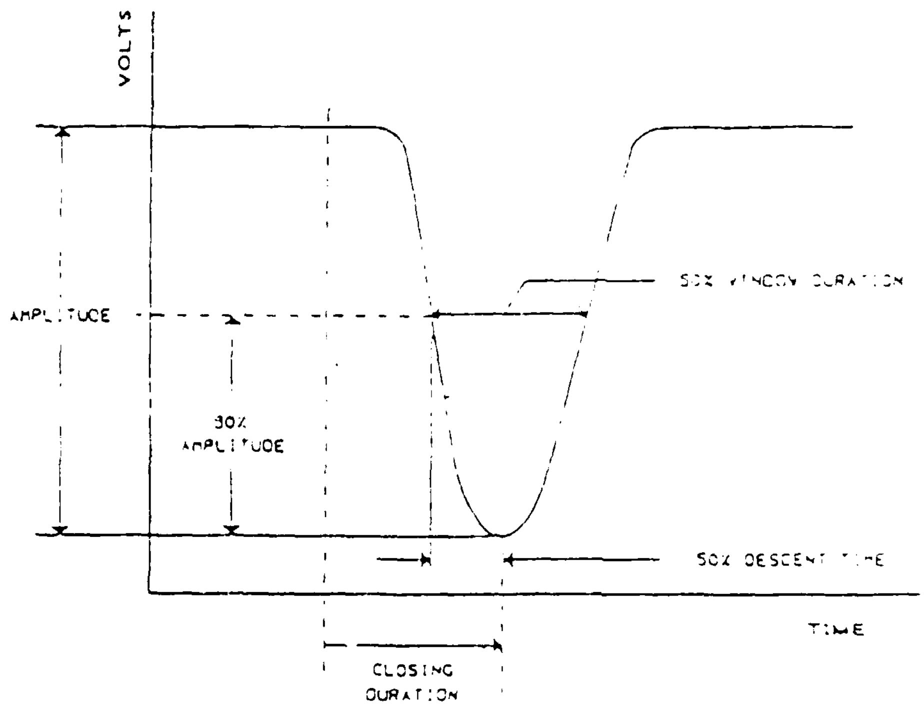

HEART RATE - BPM
90.0
90.4
91.8
93.6
95.3
"OUTER MARKER"
MIDDLE MARKER
15 SEC. ZONE
15 SEC. ZONE
15 SEC. ZONE
15 SEC. ZONE
15 SEC. ZONE
RUNWAY


MEAN HEART RATE - BPM
105
100
95
90
85
MM = MIDDLE MARKER
MM
APPROACHES

BLINK ACTIVITY

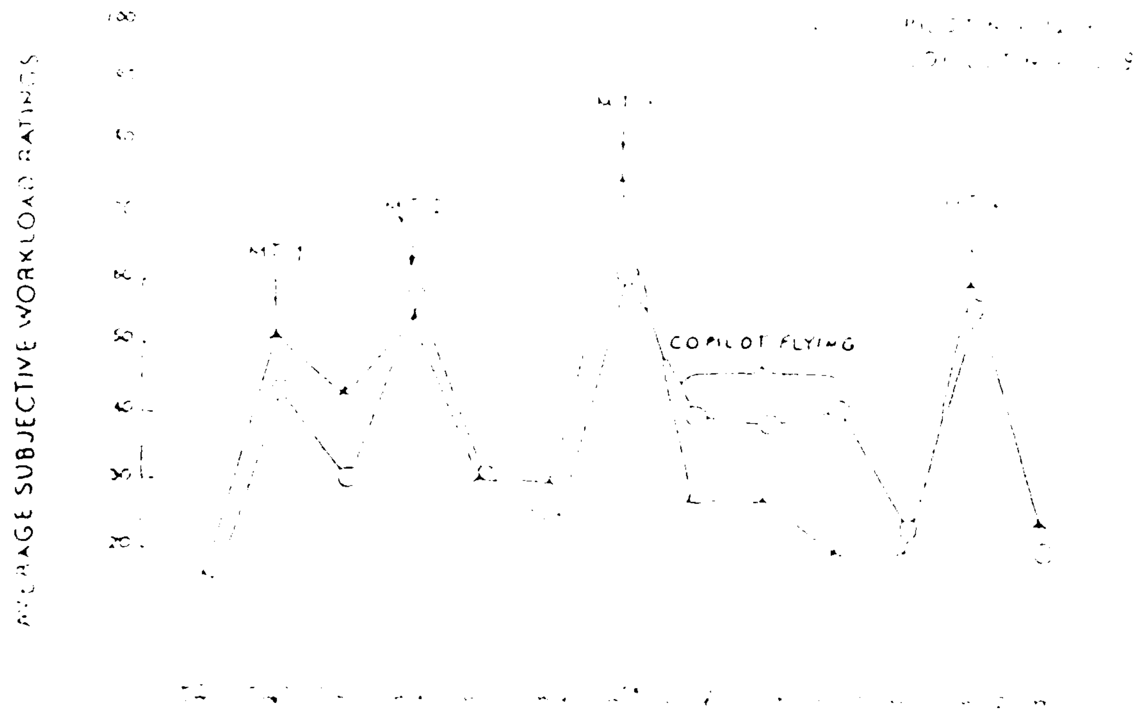
- MEASURES
  - RATE
  - DURATION
  - AMPLITUDE
- RELEVANT DATA
  - SIMULATOR
- USES
  - VISUAL ATTENTION
  - FATIGUE
- RELATIVELY EASY TO IMPLEMENT
- ACCEPTABLE TO PILOTS
- NEED MORE DATA

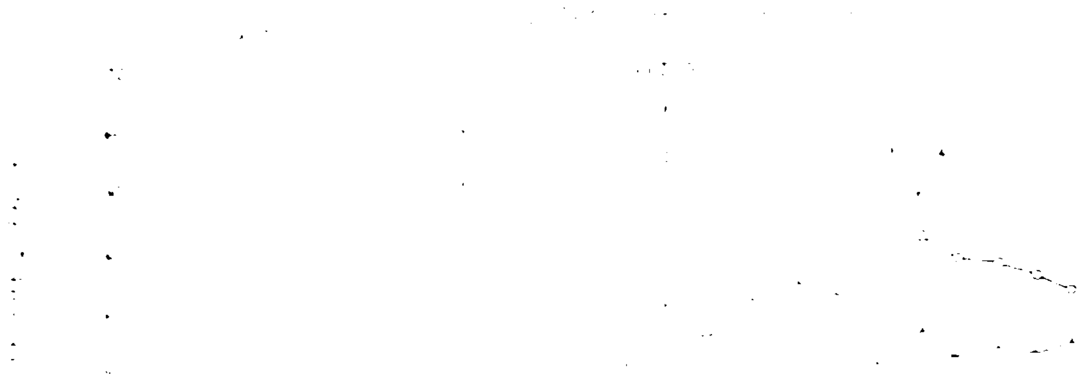
HE 87-2-9



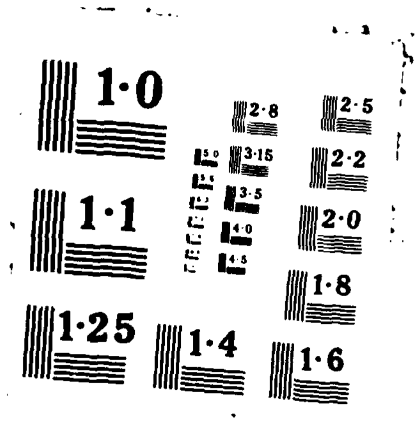
Definition of Eye Blink Parameters

SUBJECTIVE RATINGS





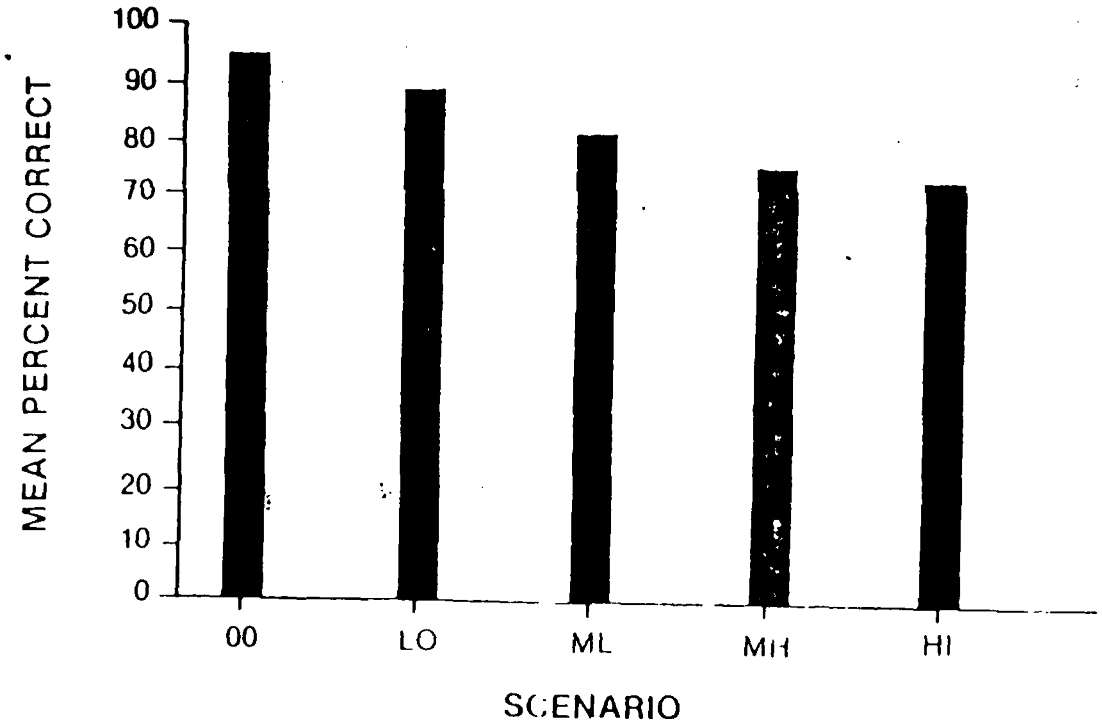
AD-A189 004 PROCEEDINGS OF THE WORKSHOP ON THE ASSESSMENT OF CREW 4/4
WORKLOAD MEASUREMEN..(U) DOUGLAS AIRCRAFT CO LONG BEACH
CA M A BIFERNO ET AL. JUN 87 AFWAL-TR-87-3043-VOL-1
UNCLASSIFIED F33615-86-C-3600 F/G 5/9 NL


1·0
2·8
2·5
3·15
2·2
3·5
1·1
4·0
2·0
4·5
1·8
1·25
1·4
1·6

EVOKED POTENTIALS

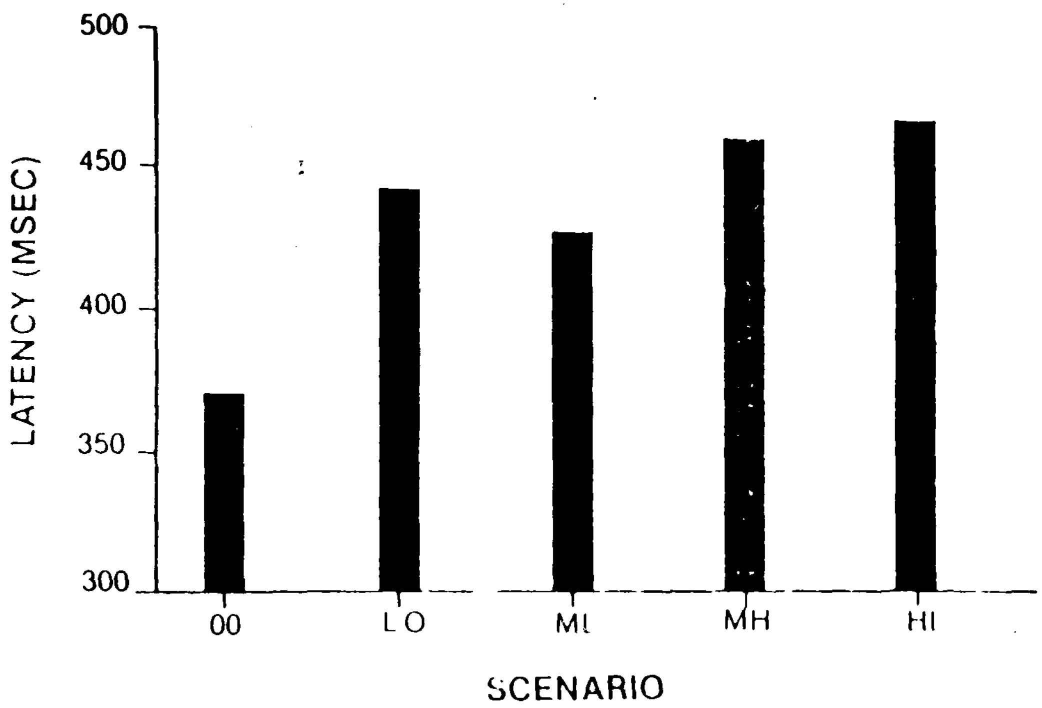
- MEASURES
  - AMPLITUDE
  - LATENCY
- RELEVANT DATA
  - LABORATORY
  - SIMULATOR
- USES
  - INFORMATION PROCESSING
  - CAPACITY
  - ATTENTION
- FAIRLY EASY TO IMPLEMENT
  - ARTIFACT PROBLEMS
- GENERALLY ACCEPTABLE
- NEED MORE DATA

% CORRECT RESPONSES
TO VISUAL RARE EVENTS



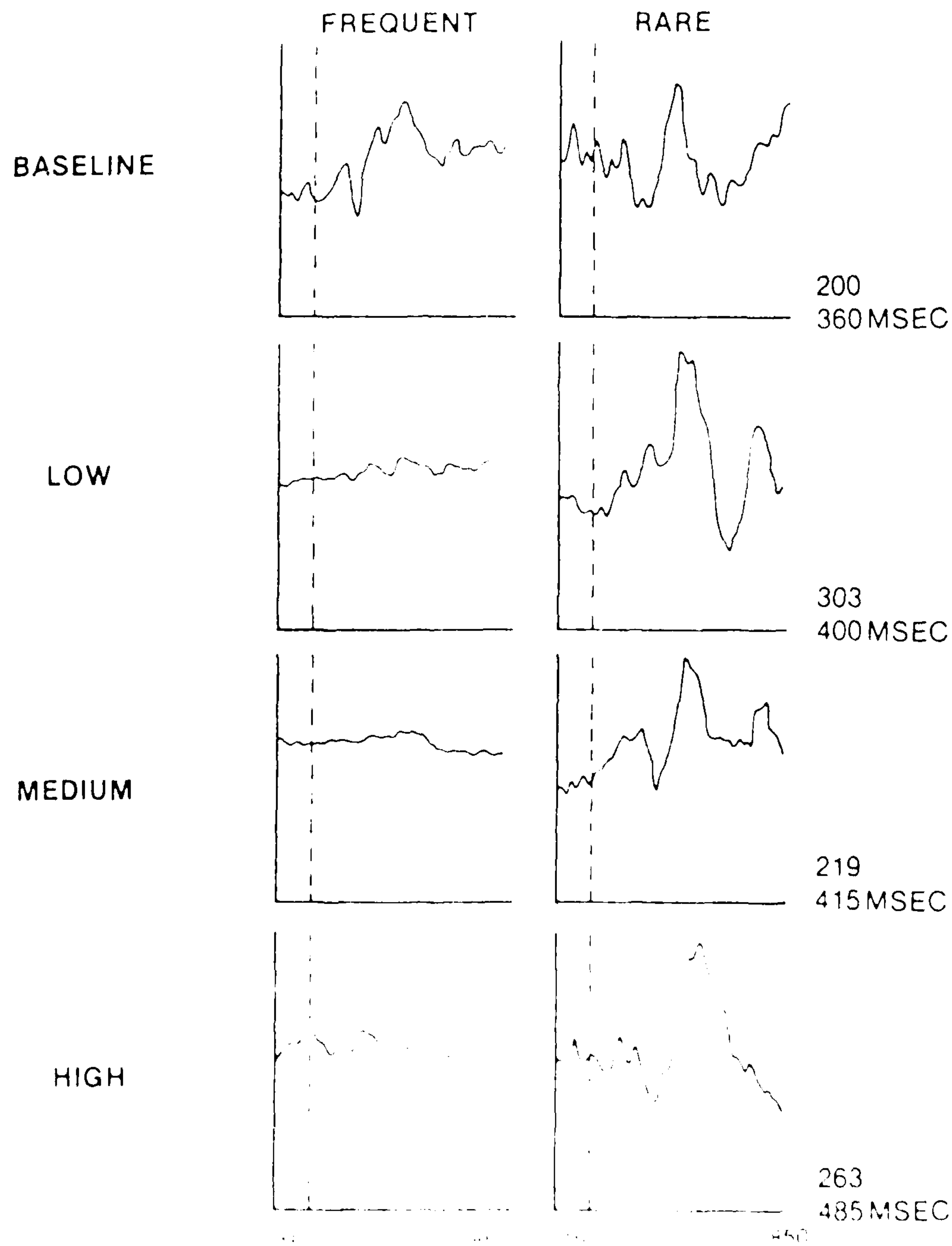
HE 86 9 102

Pz LATENCY OF P300



HE 86 9 101

VISUAL RARE EVENT EVOKED POTENTIALS



RECOMMENDED MEASURES

1. CARDIAC ACTIVITY
   — LARGEST DATA BASE
   — EASY TO USE
2. BLINK ACTIVITY
   — EASY TO USE
   — SMALL DATA BASE
3. EVOKED POTENTIALS
   — RICH POTENTIAL
   — SMALL DATA BASE
   — APPLICATION PROBLEMS

OTHER MEASURES

- PUPIL-INSTRUMENTATION
- RESPIRATION — SPEECH AND SCORING
- EEG SPECTRAL ANALYSIS — NEEDS MORE DATA
- MUSCLE ACTIVITY — MEASUREMENT & INTERPRETATION
- GALVANIC SKIN RESPONSE — APPLICATION. ETC.
- BODY FLUIDS — TIME FACTOR, INTERPRETATION AND IMPLEMENTATION

SUMMARY

- RECOMMEND CARDIAC, BLINK AND EVOKED POTENTIAL
- MUST BE COMBINED WITH PERFORMANCE AND SUBJECTIVE MEASURES
- INDIVIDUAL DIFFERENCES IMPORTANT

PHYSIOLOGICAL MEASURES

- NON-INTRUSIVE
- CONTINUOUS
- IMMEDIATE

HE 87 2 36

CANDIDATE MEASURES

- CARDIAC ACTIVITY
- BLINK ACTIVITY
- EVOKED POTENTIALS
- PUPIL DIAMETER
- RESPIRATION
- EEG ANALYSIS
- MUSCLE ACTIVITY
- BLOOD ANALYSIS
- URINE ANALYSIS

Neuropsychological Workload Test Battery (NWTB)

CENTRAL MEASURES	• TRANSIENT EVOKED RESPONSE — ODD-BALL — MEMORY SCAN — MEMORY UPDATE — SELECTIVE ATTENTION — MONITORING — BRAIN STEM — TRACKING — FLASH
	• STEADY STATE EVOKED RESPONSE — CHECKER BOARD — SINE WAVE GRATING — UNPATTERNED FIELD
PERIPHERAL MEASURES	• EYE BLINK • HEART RATE • MUSCLE
PERFORMANCE MEASURES	• REACTION TIME • ERROR SCORES

HE 85 5 32

CARDIAC ACTIVITY

- MEASURES
  - RATE
  - VARIABILITY
  - T-WAVE AMPLITUDE
- RELEVANT DATA
  - COMMERCIAL
  - MILITARY
- USES
  - STRESS
  - MENTAL ACTIVITY
- EASY TO IMPLEMENT
- GENERALLY ACCEPTABLE

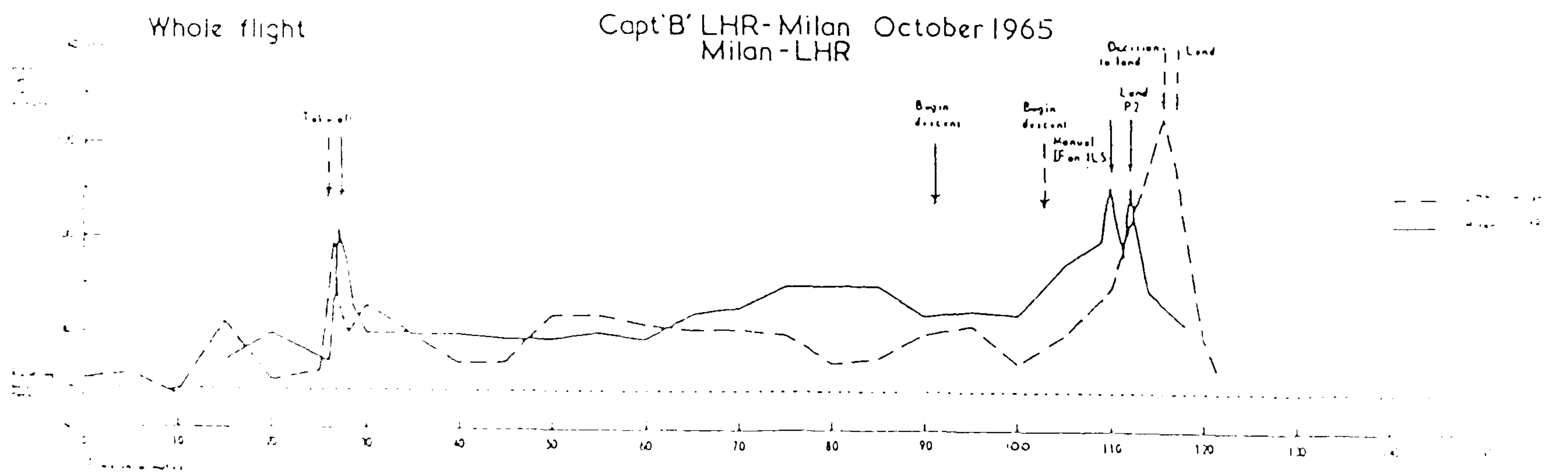
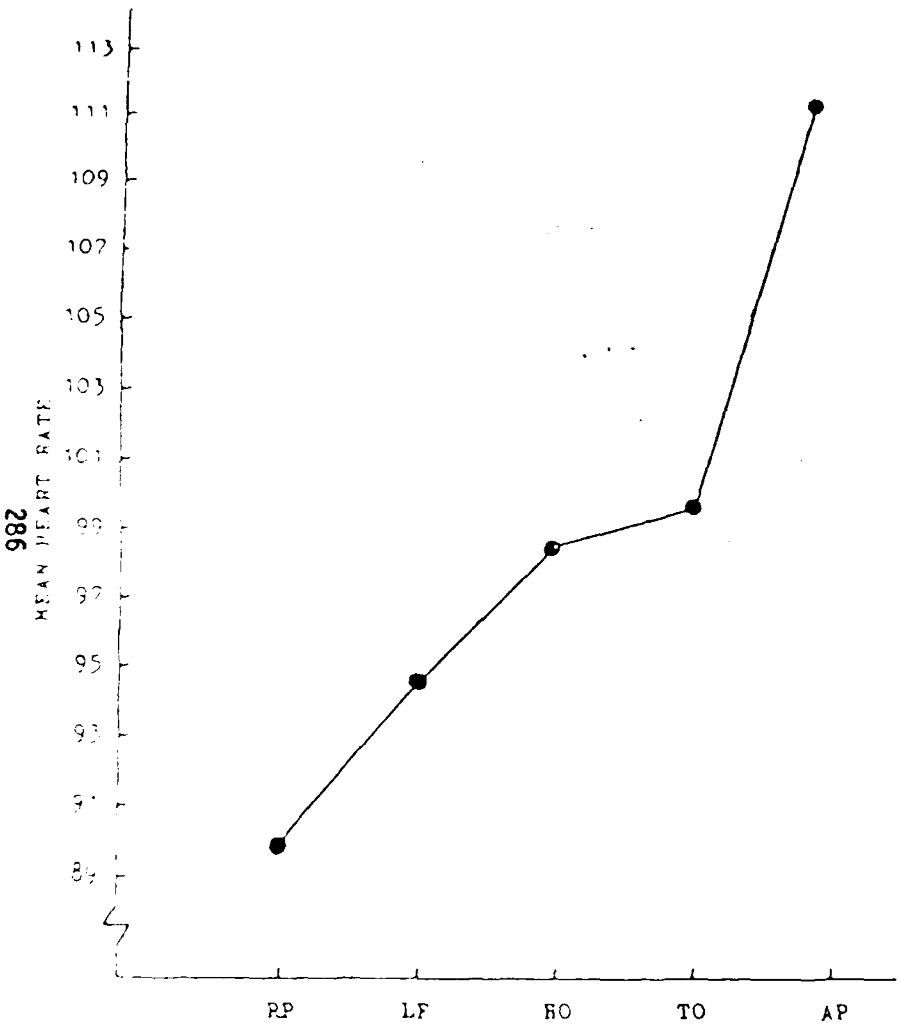


Fig. 2. Heart rate record during complete flight.

Smith, 1967



... Mean *heart rate* per min of two pilots, making nine ... on three different days in a DC-7 simulator, during ... phases of the ROT pattern. The difficulty of the ... supposed to increase from left to right.

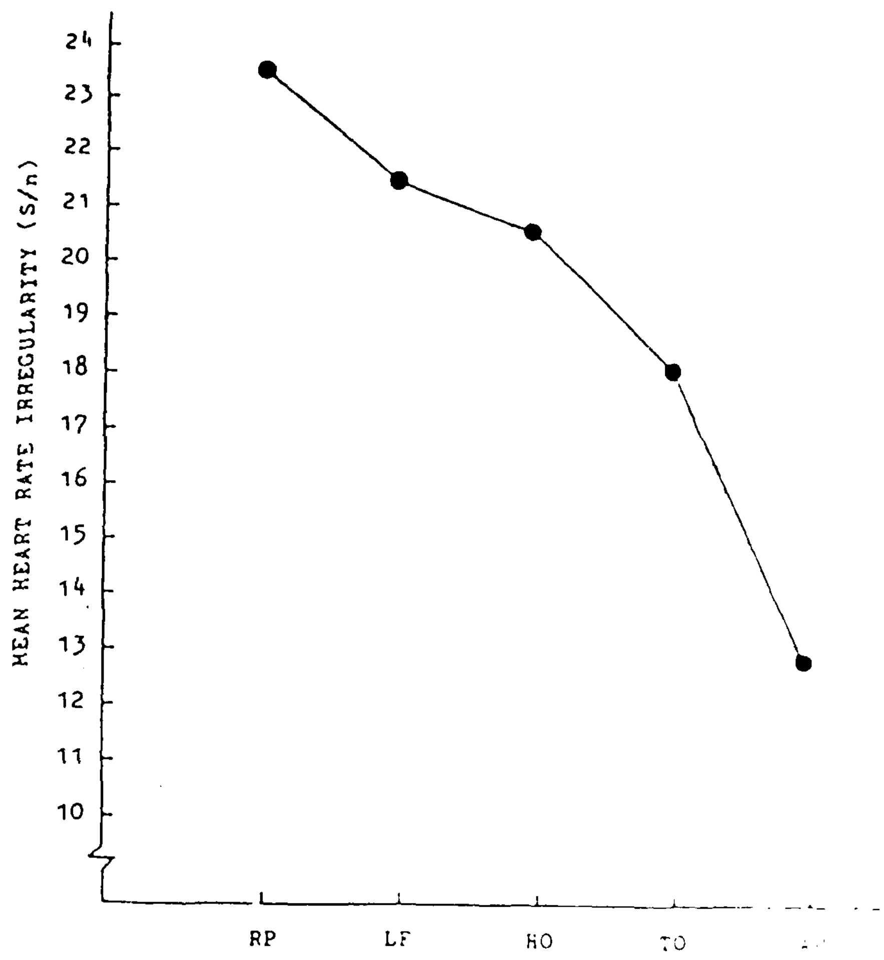


Fig. 4. Mean *heart rate irregularity* per min of two ... making nine flights each on three different days in a ... simulator, during different phases of the ROT pattern. The difficulty of the phases is supposed to increase from left to right.

Opmeer, 1973

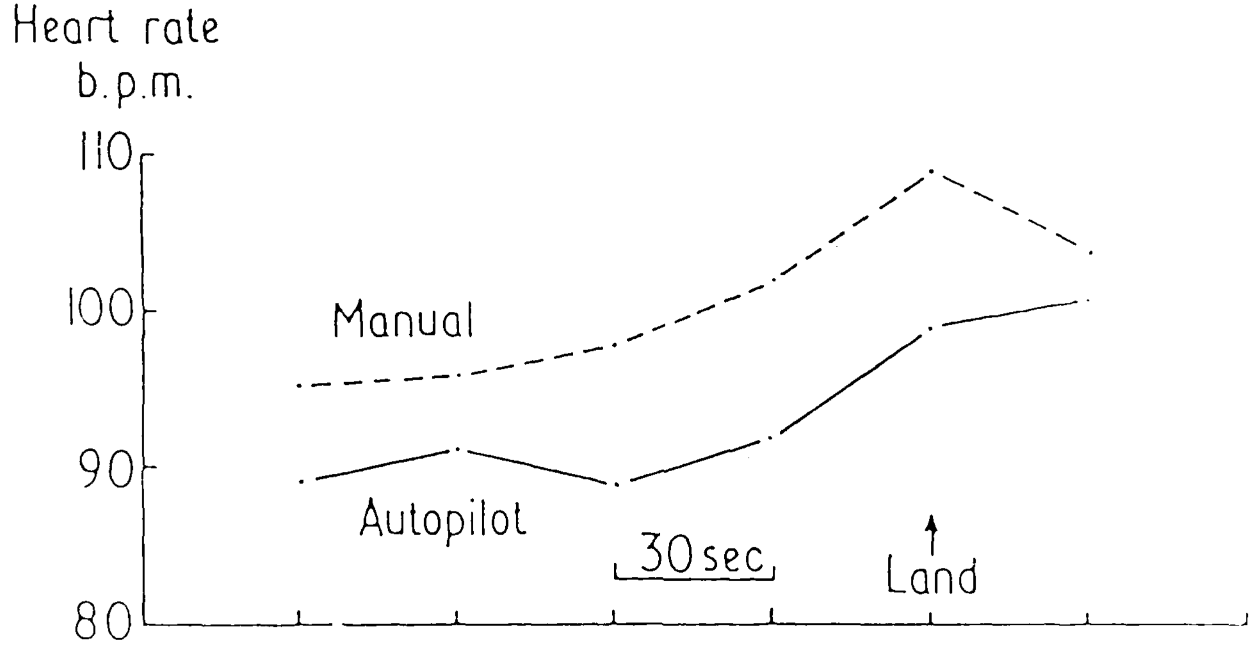
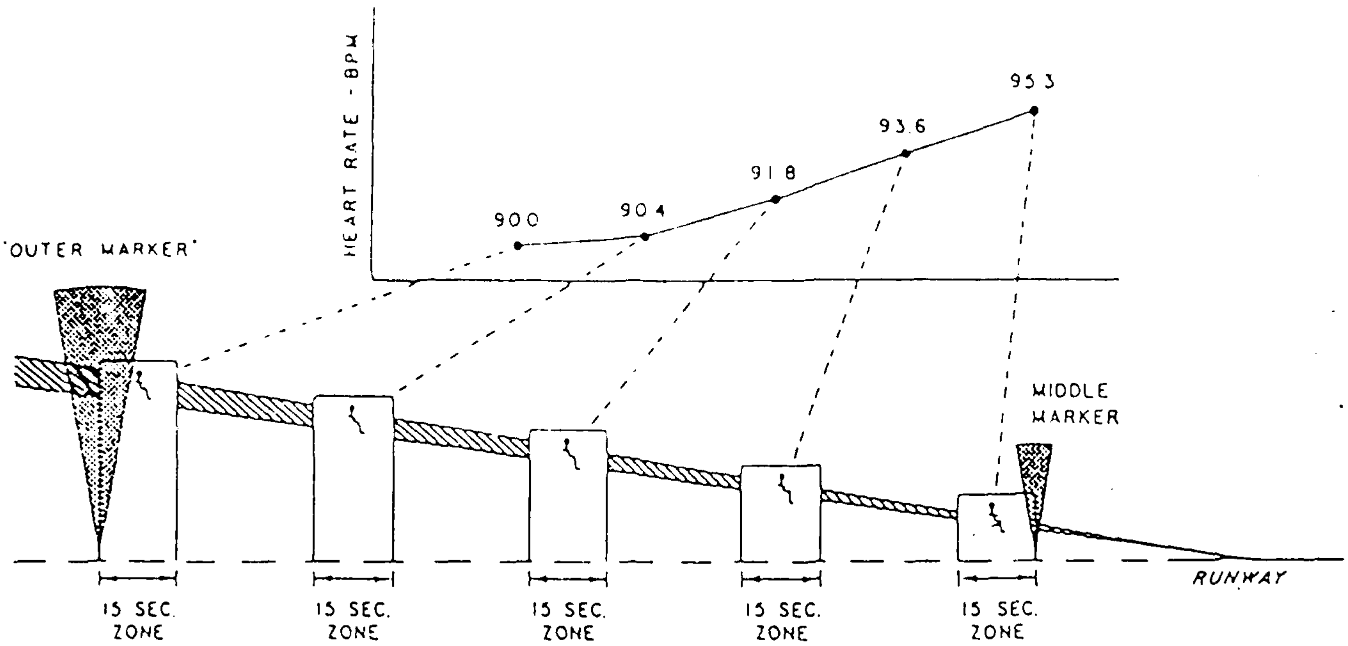


Fig. 2. Mean heart rate values (30-s epochs) for three manual and six autopilot approaches and landings during the same sortie in Comet 3B. (Pilot C)

Roscoe, 1976



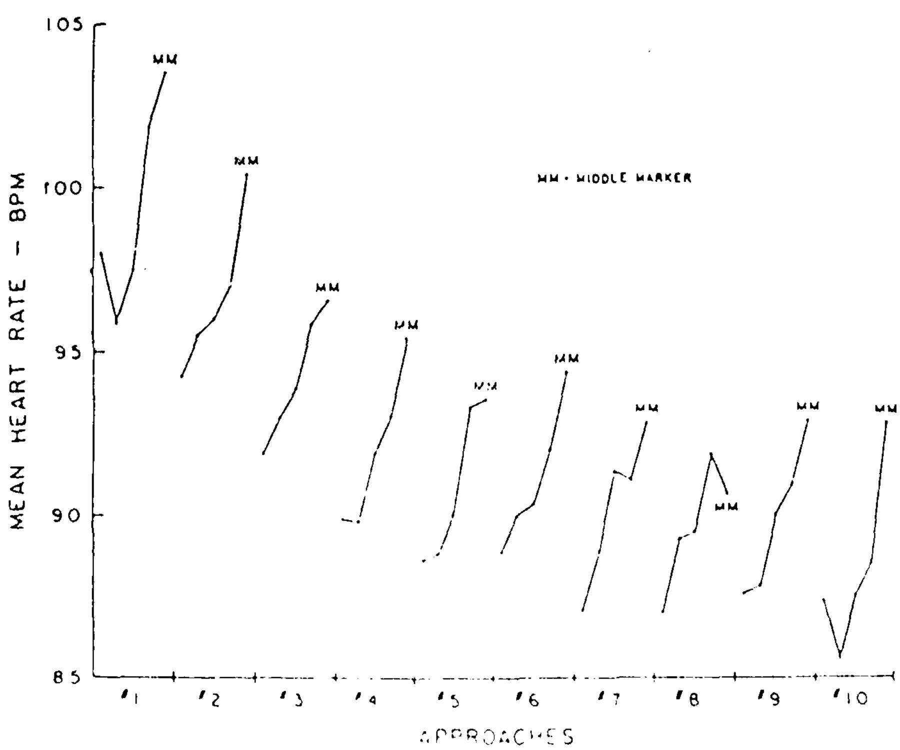
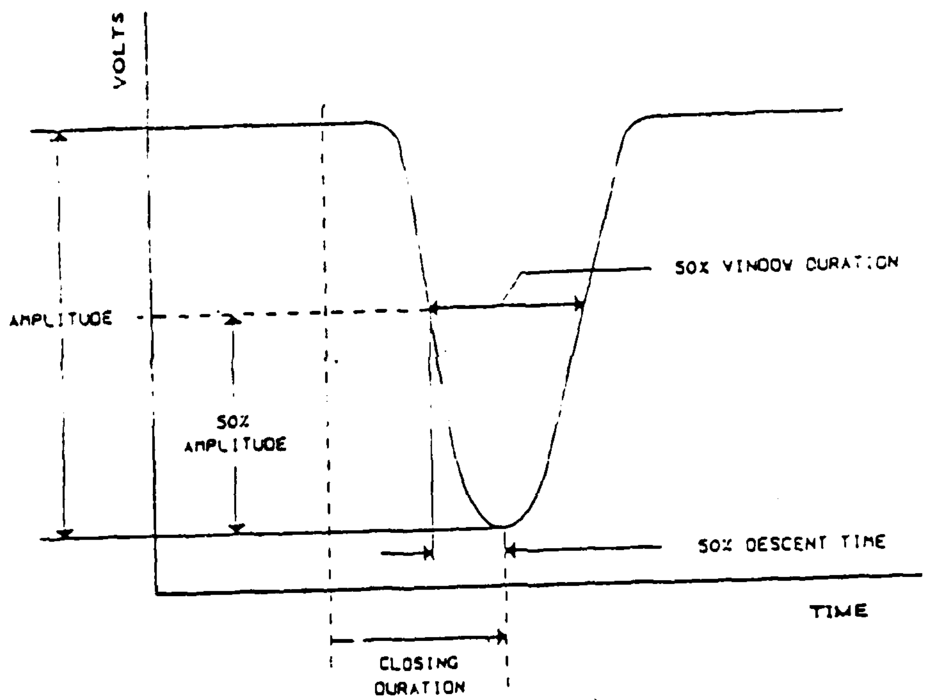


Fig. 4. Heart rate means of [illegible] pilots flying ten simulated [illegible] approaches. [illegible] curve represent from left to right [illegible] spaced along the approach path [illegible] marker.

BLINK ACTIVITY

- MEASURES
  - RATE
  - DURATION
  - AMPLITUDE
- RELEVANT DATA
  - SIMULATOR
- USES
  - VISUAL ATTENTION
  - FATIGUE
- RELATIVELY EASY TO IMPLEMENT
- ACCEPTABLE TO PILOTS
- NEED MORE DATA

HE 87 2 39



Definition of Eye Blink Parameters

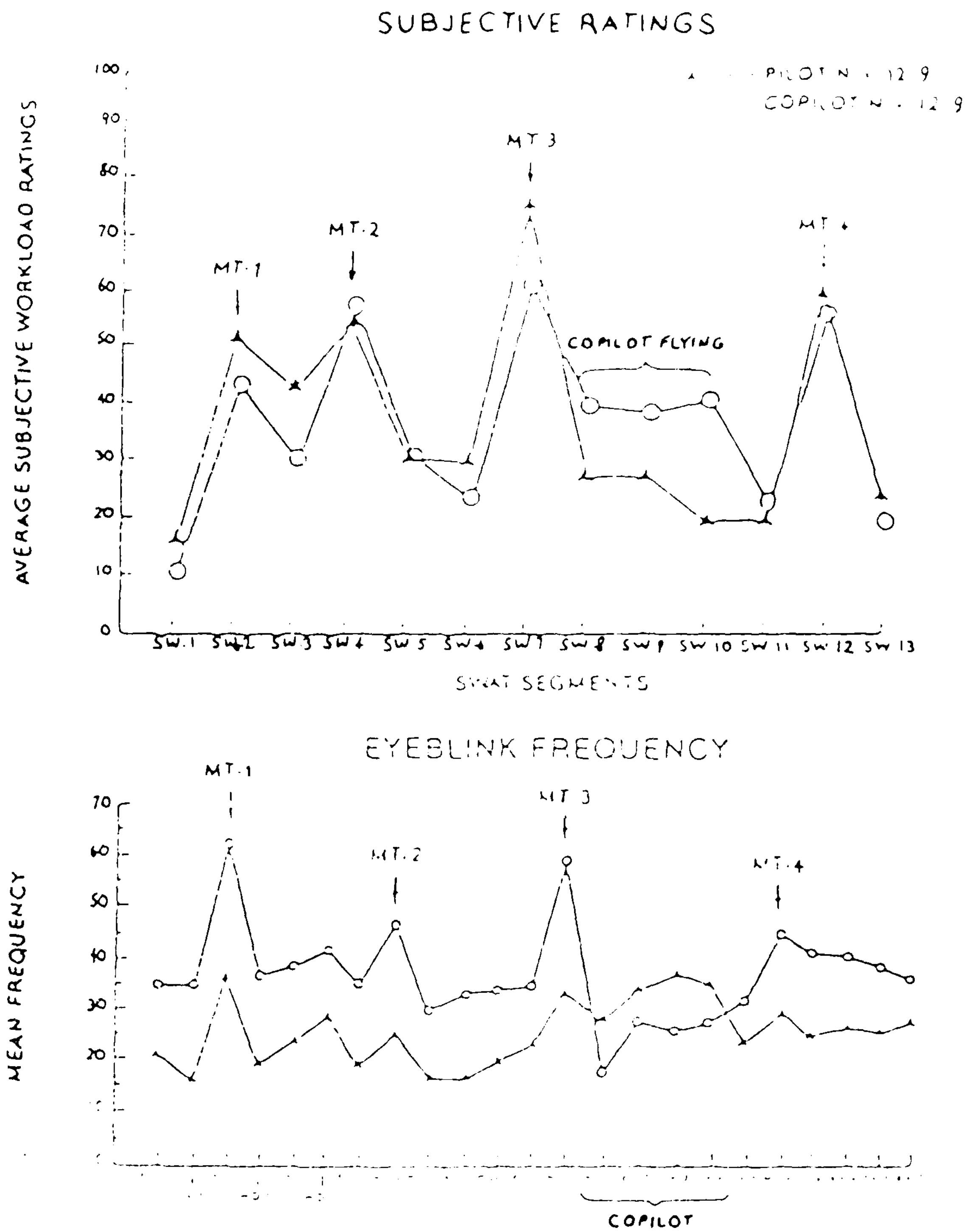


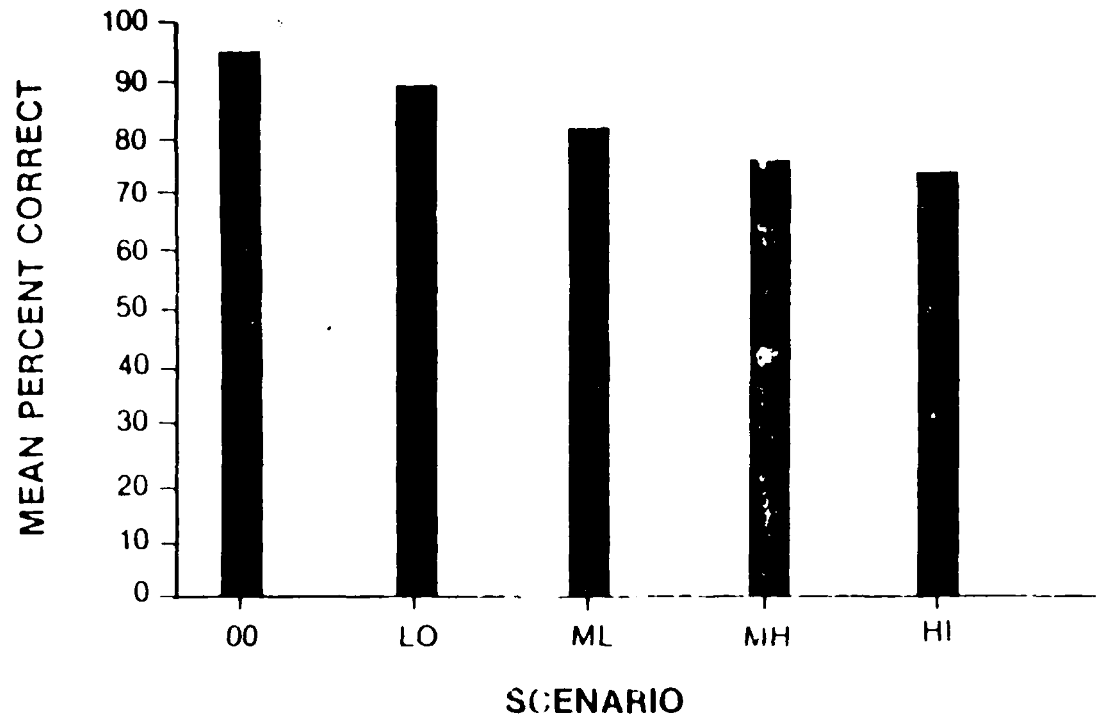
Figure 2. Workload Across Mission Segments

Skelly, 1985

EVOKED POTENTIALS

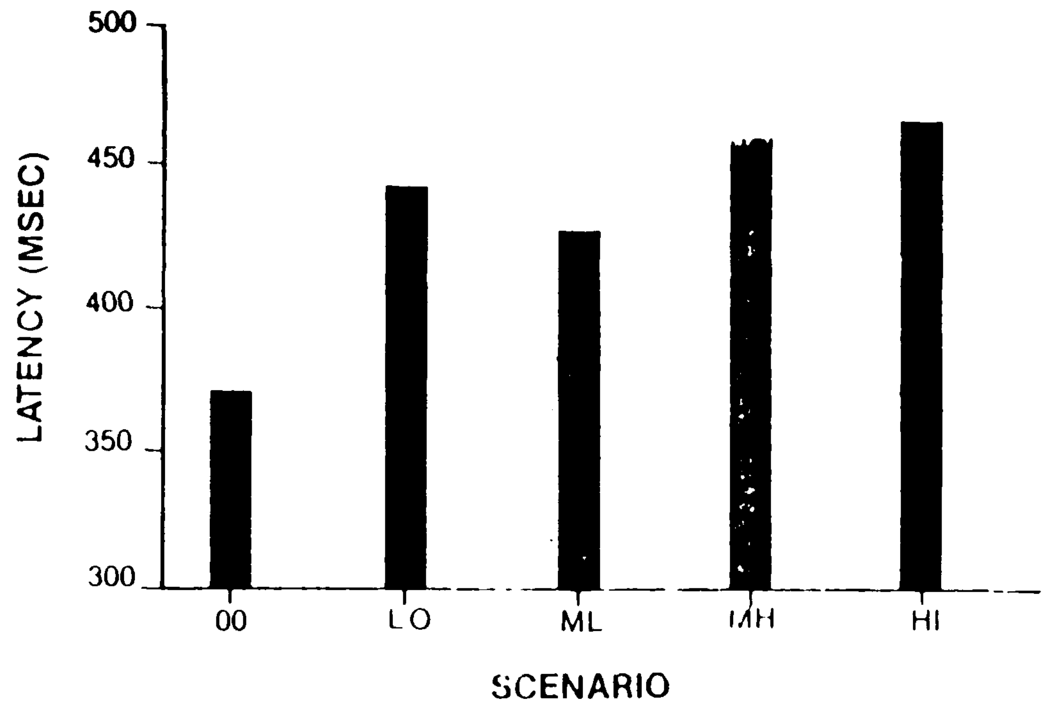
- MEASURES
  - AMPLITUDE
  - LATENCY
- RELEVANT DATA
  - LABORATORY
  - SIMULATOR
- USES
  - INFORMATION PROCESSING
  - CAPACITY
  - ATTENTION
- FAIRLY EASY TO IMPLEMENT
  - ARTIFACT PROBLEMS
- GENERALLY ACCEPTABLE
- NEED MORE DATA

% CORRECT RESPONSES
TO VISUAL RARE EVENTS



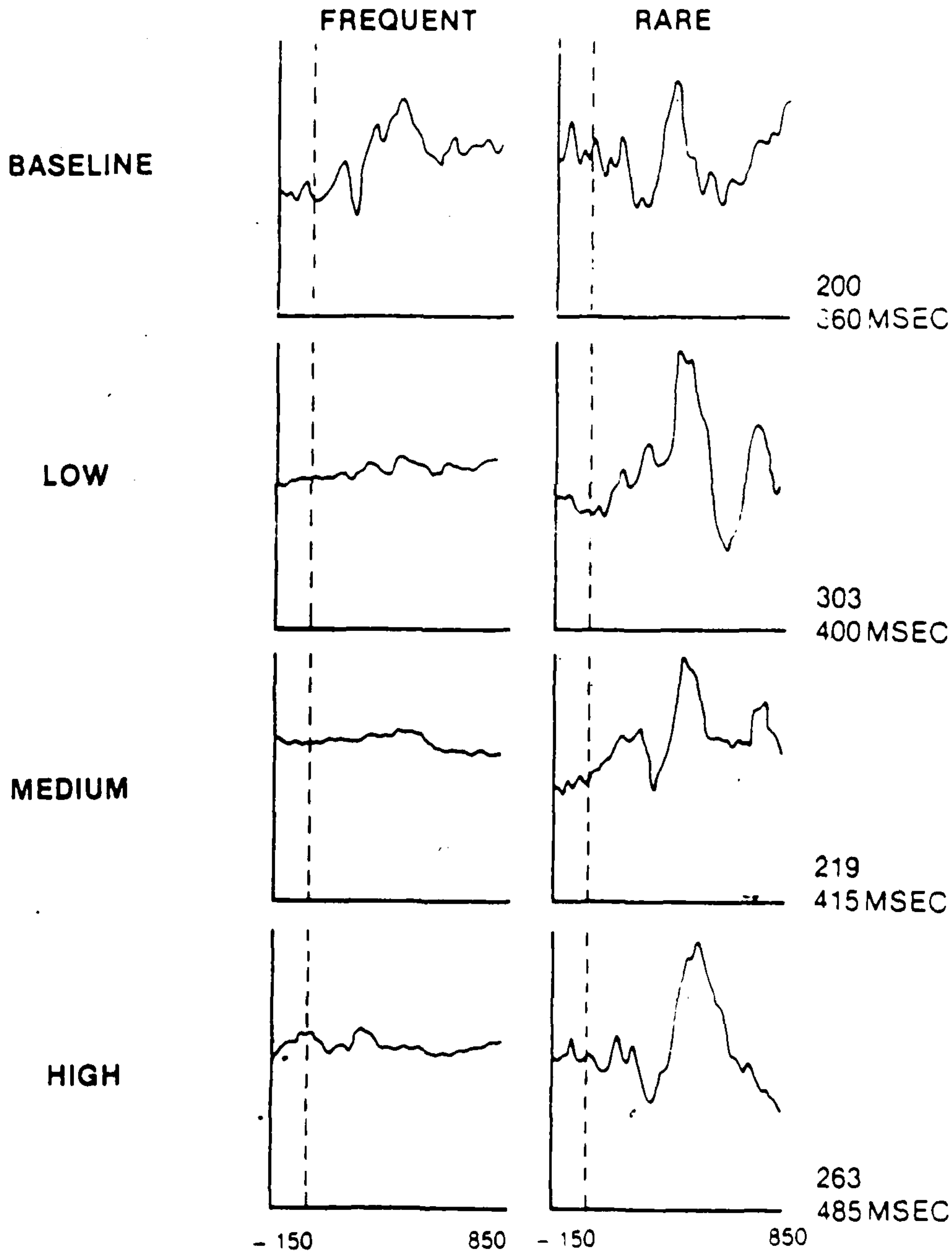
HE 86 9 102

Pz LATENCY OF P300



HE 86 9 101

VISUAL RARE EVENT EVOKED POTENTIALS



RECOMMENDED MEASURES

1. CARDIAC ACTIVITY
   - LARGEST DATA BASE
   - EASY TO USE
2. BLINK ACTIVITY
   - EASY TO USE
   - SMALL DATA BASE
3. EVOKED POTENTIALS
   - RICH POTENTIAL
   - SMALL DATA BASE
   - APPLICATION PROBLEMS

OTHER MEASURES

- PUPIL-INSTRUMENTATION
- RESPIRATION — SPEECH AND SCORING
- EEG SPECTRAL ANALYSIS — NEEDS MORE DATA
- MUSCLE ACTIVITY -- MEASUREMENT & INTERPRETATION
- GALVANIC SKIN RESPONSE -- APPLICATION, ETC.
- BODY FLUIDS — TIME FACTOR, INTERPRETATION AND IMPLEMENTATION

HE 87 2 41

SUMMARY

- RECOMMEND CARDIAC, BLINK AND EVOKED POTENTIAL
- MUST BE COMBINED WITH PERFORMANCE AND SUBJECTIVE MEASURES
- INDIVIDUAL DIFFERENCES IMPORTANT

FACT MATRIX CONSTRUCTION, CONTENT, AND EVALUATION

KENDRICK WILLIAMS
DOUGLAS AIRCRAFT COMPANY

MCDONNELL DOUGLAS

CB1335.01

OVERVIEW

SUMMARIZE FACTORS AND CONSTRAINTS OF MATRIX DESIGN AND CONSTRUCTION

SUMMARIZE CRITERIA AND CONSTRAINTS FOR INCLUSION OF ARTICLES IN FACT MATRIX

SUMMARIZE RULES AND PROCEDURES FOR EVALUATION OF FACT MATRICES BY SUBGROUPS

MCDONNELL DOUGLAS

CB1335.02

FACT MATRIX PURPOSE

PROVIDE A SUMMARY OF REFERENCES WHICH PERTAIN TO VALIDITY AND RELIABILITY OF WORKLOAD TYPES DESCRIBED IN FAR 25

MCDONNELL DOUGLAS

CB1335 03

WORKLOAD

FAR 25.1523 APPENDIX D; b.(4) THE DEGREE AND DURATION OF CONCENTRATED MENTAL AND PHYSICAL EFFORT INVOLVED IN NORMAL OPERATION AND DIAGNOSING AND COPING WITH MALFUNCTIONS AND EMERGENCIES

- MENTAL DEGREE
- MENTAL DURATION
- PHYSICAL DEGREE
- PHYSICAL DURATION

MCDONNELL DOUGLAS

CB131-65

FACT MATRIX CONSTRUCTION

- FAR 25 SCOPE LIMITED TO DEGREE AND DURATION OF MENTAL AND PHYSICAL EFFORT (FACTOR 4)
- RELIABILITY AND VALIDITY REFERENCES CROSSED WITH MEASURES
- MEASURES CATEGORIZED INTO PERFORMANCE, PHYSIOLOGICAL, AND SUBJECTIVE GROUPINGS
- APPLICABILITY WEIGHTS ASSIGNED TO MEASURES BUT NOT CROSS-REFERENCED WITH ARTICLES

MCDONNELL DOUGLAS

CB1335 04

FACT MATRIX CONTENT

MEASUREMENT DATA WHICH HAS BEEN EMPIRICALLY RELATED TO WORKLOAD

MEASUREMENT DATA CONTENT WHICH CAN BE EMPIRICALLY RELATED TO RELIABILITY AND VALIDITY

MCDONNELL DOUGLAS

CB1335 07

EMPIRICAL DATA CRITERIA

CRITERIA: MEASUREMENT DATA MUST BE STRUCTURED TO ADDRESS WORKLOAD PUBLICLY, EITHER THROUGH DEMONSTRATION OF METHOD OR THROUGH TEST OF HYPOTHESES

- MEASUREMENT DATA CLASSES:
  - CASE STUDIES (TIMELINE ANALYSIS, CERTIFICATION PROTOCOLS, ETC.)
  - EXPERIMENTS (ACTUAL FLIGHT, SIMULATIONS, APPLIED LAB, BASIC LAB)

MCDONNELL DOUGLAS

CB1335 09

WORKLOAD CRITERIA

CRITERIA: METHODOLOGY MUST TREAT WORKLOAD MEASUREMENT AS EITHER A PRIMARY OR SECONDARY TASK. IN GENERAL, THIS TREATMENT MUST BE EXPLICITLY STATED IN THE RESEARCH

MCDONNELL DOUGLAS

CB1335 10

RELIABILITY CRITERIA

CRITERIA: MATERIAL MUST EITHER ADDRESS RELIABILITY THROUGH HYPOTHESIS AND EXPLICIT STATISTICAL EVALUATION

OR

DEMONSTRATE RELIABILITY THROUGH REPEATED EXPERIMENTAL MANIPULATION WHICH SATISFIES ONE OF THE RELIABILITY CONTENT AREAS (INTER-RATER, ETC.)

MCDONNELL DOUGLAS

CB1335 11

VALIDITY CRITERIA

CRITERIA: RELATE EXPERIMENTAL HYPOTHESES AND/OR DATA TO WORKLOAD KNOWLEDGE BASE

OR

ADDRESS MEASUREMENT APPLICATION AND PERFORMANCE, RE: OTHER "ACCEPTABLE" WORKLOAD MEASURES

OR

TEST WORKLOAD DOMAIN HYPOTHESES AGAINST A THEORETICAL MODEL OF PERFORMANCE OR WORKLOAD

MCDONNELL DOUGLAS

CB1335 12

FACT MATRIX EVALUATION

REVIEW MATRICES FOR CONTENT COMPLETENESS

ANALYZE CURRENT CONTENT FOR CORRECTNESS

SUBMIT ADDITIONAL CONTENT FOR INCLUSION

MCDONNELL DOUGLAS

CB1335.13

FACT MATRIX COMPLETENESS

REVIEW COMPLETENESS OF RELIABILITY/ VALIDITY DEFINITIONS

REVIEW MATRIX CONTENT BY MEASURE TO ASSURE ALL ARTICLES WHICH ADDRESS RELIABILITY AND VALIDITY OF MEASURE ARE INCLUDED

JUSTIFY INCLUSION OF ADDITIONAL ARTICLES IN WRITING

GATHER ADDITIONAL ARTICLES WHICH HAVE BEEN JUSTIFIED FOR FACT MATRIX INCORPORATION

MCDONNELL DOUGLAS

CB1335 14

FACT MATRIX CORRECTNESS

EVALUATE ARTICLES BY MEASURE TO ASSURE REFERENCES ARE APPROPRIATELY ASSIGNED

JUSTIFY, IN WRITING, ARTICLES WHICH ARE IMPROPERLY ASSIGNED

PREPARE LIST OF ARTICLES TO BE DELETED FROM FACT MATRIX

MCDONNELL DOUGLAS

CB1335.1:

Subjective Group Meeting - 2/25/87

Synopsis: NASA TLX and SWAT were the measurement techniques recommended for use in the Part Task testing. There are enough references and data collections using these measures to demonstrate their validity and reliability. The Modified Cooper-Harper, a unidimensional scale, was discussed but not recommended as it gives a "number" but does not allow you to work backward to how the number was developed and thus, offers no diagnosticity.

Meeting Summary

Danny Gopher said the goal of the meeting was to come up with recommendations for workload assessment methodology in certification of aircraft. We must review the problem of certification as well as the process and identify key issues in certification where our methodology can help.

Del Fadden explained the Boeing certification process and showed the Pilot Subjective Evaluation (PSE) form that is used by Boeing during the certification program. In this program the pilot is asked to compare the new aircraft with a known reference aircraft. If a pilot says a task is more difficult, then he must discuss why it is more difficult.

Del pointed out that before certification, in simulation, most of the cockpit design work has been part task. A lot of design is done even before the simulator is available. An airplane evolves - other planes are used for reference.

Jean-Jacques Speyer said Airbus takes a unidimensional view. They ask for lots of ratings, actually take ratings continuously on line. He proposed that all ratings together give a microscopic evaluation as to what is going on. The Air Worthiness observer also gives a subjective rating. If the observer sees that the pilot has not made a rating at some particular time, he can ask the pilot to give ratings at that point. (Instrument set up - observer gives crew member a light and pilot is invited to give a rating at that time). Ratings are not pass-fail, ratings are given on an absolute scale.

Group discussion followed on the two techniques. In summation, Boeings' method is a comparative measure. Airbus uses an absolute technique. Boeing takes pilot ratings post flight. Airbus takes ratings in-flight and at the end of the flight pilots are given a long questionaire - 150 questions.

Gopher said we must define the problem. Why are subjective ratings important? What does pilot feel about the task he must perform? Is he comfortable with it? All participants

agree that they use pilot subjective opinion in one form or another. Gopher felt it would be useful if we could standardize and use the same scale in all nations so we can learn from each other

In the discussion that followed on standardization the following points were made:

1. We need experienced pilots to make judgements.
2. We need to have some comparative point of reference.
3. Ratings should be made under some specified optimum conditions.
4. Raters must be trained to rate.
5. Everyone must agree on the dimensions and terms of workload so results will be communicable.

The point was made that if the same method was used in design and certification, then there would be no surprises, although the method would be used in certification in a more limited sense. Del Fadden pointed out that this is not always possible, for example, for the design of much of the 7J7 there is no reference airplane.

Danny Gophers opinion on selection of rating scales is to use NASA TLX and SWAT. He said they are the most viable and best documented. He said the Modified Cooper-Harper is "dead", not being used and is based on old research. He would like to see both NASA TLX and SWAT used, at least to start.

Modified Cooper Harper

Sandy Hart looked at the Modified Cooper-Harper when she began to develop a scale. In this scale the raters move through a series of decisions, these decisions determine whether the rater ends up at the top, middle, or bottom of the scale. You can't take the final number and work backwards to determine how the number was developed. NASA wanted a diagnostic scale that would allow them to work backwards.

NASA TLX

Subjects were asked to give their definition of workload before they started the tests to reduce intersubject variability. During the tests subjects were asked what they felt caused the primary source of workload, then that task was given more weight. This works well for simple tasks, but weights don't have as much effect in complex tasks. It was felt that weights should be used for both simple and complex t[illegible], however

It was pointed out that the FAA is interested in the 6 basic workload functions listed in the FARs so perhaps a scale

should be developed to evaluate these 6 functions. These functions are; flight path control, collision avoidance, navigation, communication, operation and monitoring of aircraft engines and systems, and command decisions. It was suggested that the FAA could be asked to weight these functions so the manufacturers would have a standard to work against.

SWAT

Gary Reid said he and his colleagues did not set out to develop a subjective scale. They were asked as workload experts how to measure workload.

They started out with the Simpson-Sheridan scale. The disadvantage of this scale is that it takes a lot of training to properly use its decision tree rating scale. Next they then took a three dimensional scale, regressions approach, looking for the smallest set of dimensions that would tell them the most. They were driven by practical considerations such as cost. For example, is increased sensitivity worth the cost? This scale fits their needs. They can gather ratings with some diagnosticity. In their ratings, they feel that if they hadn't had observers, they couldn't tell why workload is higher in some instances.

SWAT covers 4 of the 6 functions listed in the FARs. Performance and physical demand are not covered. Gary said that if they need performance they measure it. Measuring mental workload is important. Physical is not as important as it once was.

Discussion followed on the importance of getting down to a "one number" assessment, is the extra effort to get there really necessary? It was felt by some "one number" provides structure and information and is a valuable adjunct to what is going on.

Danny Gopher said that the charge of the subjective group was to evaluate measures for use in the simulation study. It has been concluded that it is important to use subjective measures. He reiterated that all manufacturers use subjective measures in one form or another. This group strongly recommends the use of subjective measures. It is reasonable and advisable to find a systematic way to elicit these opinions. He feels that standardization is one approach. How to get the information from the pilots is the question?

Both NASA TLX and SWAT give global assessments of tasks as a whole. There are enough references and data collections using them to provide valid and reliable results. The Airbus on-line measurement model is different because one must design situations so a single number will be meaningful

and one must have many such numbers. It is a costly and demanding process and is not well documented. Gopher would like Airbus to try NASA TLX and SWAT along with their method to see if the methods agree.

Discussion followed on the importance of using both NASA TLX and SWAT. The following points were made;

1. The way in which you communicate with the pilot is different.

2. There are differences in the weighting systems of the two scales

3. It is important to generate more than just a single measure of workload.

The concept was discussed that assessment is a package, not a scale. Raters must be qualified and pilots must be trained. NASA TLX and SWAT are both used by in flight, as often as possible. Ratings are taken as soon as possible. Both Sandy Hart and Gary Reid stated that if data is collected in flight, they are more comfortable with it. Results of tests, however, do not back up feelings that in-flight measurement is better. Gopher felt that if we can get data post-flight with no change in results, that would be an important finding.

Much discussion followed on what number is good enough. Should some subjective measures be taken in-flight and some post-flight? If the measures are taken in-flight, are they intrusive? It was the feeling of the group that measures don't always have to be intrusive, for example, after landing you can gather ratings on outer, middle, inner marker and landing.

Danny Gopher said he would summarize the subjective group meeting results in the general meeting for 15 minutes, then allow Sandy Hart and Gary Reid to give summaries since they don't agree with Danny on many points. Sandy and Gary said they gave their presentations yesterday. Sandy stated that Gophers opinion does not accurately reflect that of the group.

MINUTES

Performance Group Meeting - 2/25/87

Summary -

The performance group discussed the cost of measuring workload for behavioral, physiological, and subjective measures. Much support was given for including a Sternberg Task in the simulation test battery.

Several valid and reliable measures and implementation requirements were agreed upon. The measures suggested were: Sternberg task, critical tracking, choice reaction time, mental arithmetic, and embedded communication.

The Cost of Measuring Workload -

Performance measures (procedural errors): Measure impairments to performance.

Physiological measures: Measure psychosomatic effects of stress and occupational diseases.

Subjective measures: Measure conscious experience, estimate the ability to cope with goals and achieve criteria, and are sensitive to one general "work of intentions". This influences performance on the very general level. The costs or possible effects of subjective measures are that misjudgements may affect selection of goals and criteria, may affect motivation, may affect risk taking behavior.

Types of Certification -

The group established that certification varies depending on what organization is doing the certification.

British: Look at the relationship between subjective, physiological, and behavioral. A study by Jean Jacque Speyer was mentioned testing the possibility of using control reversals during approach as a measure.

Boeing: Weather is a random event. A simulator can impose whatever weather desired. Certification uses a performance margin. Failures, weather, etcetra are added. Frequency of occurrance can be manipulated. Performance margin can be large or small. Uses traditional timeline analysis.

Implementation Requirements for User Acceptance -

The group developed the following requirements to insure user acceptance:

1 - The measure must be non intrusive.

2 - The measure must conform to all safety standards.

3 - The tasks must be within the realm of "normal" methods.

4 - The measure must not lower crew self image.

5 - The measure must be non career threatening.

Valid and Reliable Measures -

The group suggested and agreed upon the value of the following measures:

1 - Sternberg (auditory/visual) can be highly intrusive. Must make efforts to move toward "real" or "normal" tasks.

2 - Critical tracking (psychomotor) can be too complicated. This measure was considered of boarderline acceptability but the group voted to keep it as it can be useful.

3 - Choice reaction time (visual).

4 - Mental arithmetic (auditory/visual) allows for much flexability. Can be embedded and made "realistic" to piloting tasks.

5 - Embedded communications (auditory) "normal" secondary tasks.

Problems and Considerations -

The group raised the following problems and considerations:

1 - There may be a problem with comparing a new craft's performance with a "reliable and safe" old craft's performance. The comparison will not always be reliable. New crafts are often too different from old crafts. This can yield misleading results.

2 - Must consider that the design stage implements engineering test pilots, the testing stage implements line pilots, and the certification stage implements engineering test pilots.

3 - Theory must guide implementation.

4 - Must consider the training on, and difficulty of, tasks (data vs. resource limit).

5 - Must consider fatigue from introducing the tasks in terms of adding energy exp[illegible]iture to the crew especially at work underload.

Sternberg Task as a Measure -

The group cited several workload measurement studies looking at handling, displays, and crew coordination aspects of flight. All studies used a Sternberg task.

1 - 1982, this study evaluated 2 HUD display formats. The measure used was a visual Sternberg task. Pilots flew ILS. While the experiment was in progress, the experimentors recorded several verbal responses such as whistling and short comments about the task.

2 - This study varied frequency of input during a terrain following flight director profile. An adaptive visual and auditory Sternberg task was used. A top level of error was preset. The study found that the visual Sternberg task discriminated between different levels of workload while the auditory did not.

3 - This study varied visual disorientation with a Malcom horizon. The measure used was a Sternberg letter presentation task. Results were mixed.

4 - Dunn, 1985, Helicopter Proceedings Conference. This study tested kenisthetic displays. Pilots "flew by the feel" of the instruments.

5 - This study tested a cross coupled instability tracking system. The measure used was a Sternberg variable pitch task where the instability level was varied.

After reviewing these studies, the group agreed the Sternberg task should be among the measures tested during simulation.

Questions Raised by the Group -

1 - How to address user popularity? Aspects of a potential measure may appeal to large air line manufacturers, but not to small general aviation manufacturers or vice versa.

2 - How to assess dimensionality? What is being measured?

3 - How to address the demonstration of worst case? What are the probabilities?

4 - How to assess practicality? We must consider cost, degree of simulation, and flight testing.

5 - How to address methods? We must determine where, when and how they are relavent. We must decide whether or not to test the limits.

6 - How to tailor the measure to each situation in which it is used?

7 - How to do a comparison of techniques?

Performance

- Certification Process
  - Phases
- Primary Task
  - Control Activities
  - Errors
    - Performance Margin

Secondary Tasks

Part Task → Simulator → Flight Test

- Sternberg (Auditory, Visual)
- Choice Reaction Time
- Mental Arithmetic
- Critical Tracking

. . . .

- Embedded (communication)

Caveats:

- Theory in task application
- Crew vs. individual

[illegible]

- Background
  - Certification process - phases

- Primary Task
  - control activities (reversals)
  - errors (procedural WRT system failures)
    - Performance Margin - Total perf, [illegible] etc [illegible] [illegible] that [illegible]
- ~~[illegible]~~

- Secondary Tasks

Part Task — Simulator - Flight Test

STERNBERG - Call Sign* example

Implementation fidelity

Tasks

CAVEATS &
- Theory in application of tasks [methodol
- Crew vs Individual

~~John Stern~~

USAF/FAA REVIEW OF WORKLOAD MEASUREMENT METHODS

SUMMARY REPORT:

Physiological Measures Group

Prepared by:

John A. Stern, Ph.D.
Washington University Behavior Research Laboratory
at Malcolm Bliss Mental Health Center
1420 Grattan Street
St. Louis, Missouri 63104

Recommendation for immediate implementation

1.0 Constraints

1.1 Measure must be usable in aircraft cockpit and simulator.

1.2 Measurement procedure must not impose secondary task performance requirements on subject.

2.0 Recommended measures

2.1 Heart rate and heart rate variability (HR).

2.2 EOG: eye blink and eye movements.

2.3 Perhaps aural canal temperature.

3.0 Rationale for choice of measures

3.1 HR and HR variability

Reports by Mulder and Moray dealing with HR variability as measure of mental effort and comments by Alan Roscoe concerning HR.

3.1.1 HR variability as defined by Mulder

- Abstracts interbeat interval (IBI) for fixed number of IBI's (IBI accurate to ± 5 msec). Moray uses 256 to compute.
- Conduct Lagrange interpolation for every 500 msec.
- Autocorrelate these equidistant values with maximum lag of 10% of computed values.
- Fourier transform autocorrelation functions.
- Smooth raw spectral densities with Hamming window.
- Derive power of five spectral points (.06, .08, .10, .12, .14 Hz).
- Derive natural log of power of the five points.
- Activity in .06-.14 Hz is sensitive to "mental effort" manipulation. Greater energy reflects relaxation, lower energy mental effort.

3.1.2 Comments about procedure from principal advocate(s)

- Morey reported that a variant of this procedure is currently in use in his laboratory and works to index mental work load in a variety of situations. He also reported that although he used the Mulder algorithm, he could not replicate Mulder results in earlier investigations. The cause of this inability to replicate is a mystery. Other researchers have experienced and reported the negative results.

- Alan Roscoe reviewed the utility of HR as a measure of "arousal," "attention," or "alertness." This measure should be useful in long duration flights. In the present context (30 minute flights), it may not hold much promise, although one would expect that the two failure conditions may well be associated with immediate increases in HR. What will be difficult to determine is whether such increases are associated with the psychological threat produced by the equipment failure or an increase in motor activity associated with dealing with the problem.

  In any case, HR variability cannot be measured without acquiring HR information, therefore, HR remains as a measure in our proposed "battery."

3.1.3 Potential problem with respect to part-task simulation run as described.

- Some of the segments are only 1 or 1.5 min in duration. This may be too short a time period to obtain reliable spectral information.

3.1.4 Solutions

- Don't analyze data for segments less than two minutes (or 250 seconds) with Mulder procedure.

- Combine data from equal work load segments?

- Sample data for longer periods, if at all possible.

- Await Moray's "on line" filter technique to analyze data.

- Try $\hat{v}$ technique for Vagal tone - Porges Black box.

3.1.5 The recording of heart rate or period poses no major problem. Such measures have been successfully recorded both in the simulator and under flight conditions.

3.2 EOG - Eye blink and eye movements

- Recommendations based on results of AAMRL/HEG (Wilson), USAFSAM/VNE (Miller), and WUBRL/MBMHC (Stern). Measure has been used in laboratory and simulator, and is currently being used for in-flight monitoring (AAMRL).

3.2.1 Blink measures utilized

- Blink rate obtained results: differences in rate between pilot in command of aircraft and second in command.

- Increase in blink rate as a function of time-on-task have been demonstrated in a variety of conditions.

- Inhibition of blinking is related to visual task difficulty.

- Blink amplitude: not very discriminating for effort at hand. USAFSAM data on smaller amplitude blinks being associated with poorer performance (probably function of partial lid closure as sleep deprived subjects perform long duration (4 1/2 hours) in simulator (Morris data).

- Blink closing and/or closure duration (50% window): Closing duration is defined as time from blink initiation to peak closure. Closure duration is defined as time between the blink entering a window defined by half amplitude of blink and exiting that window. Closure duration has been shown to be sensitive to visual task demands (work load), with shorter closure duration blinks associated with more demanding tasks (demonstrated in laboratory and simulation settings).

3.2.2 Comments about procedures from principal advocates (Wilson, Stern)

- Data can be collected in simulator and in flight. Appears to be sensitive to the type of work load manipulations proposed in McDonnell-Douglas/Boeing joint effort.

3.2.3 Potential problems

- Duration over which data is sampled for various work load levels is at the lower end of acceptability. One minute of data is too short to produce reliable rate data; may be acceptable for other blink measures.

- Since eye position information will be advocated as a promising measure for future implementation, I (not the working group) would like to recommend the recording of horizontal eye movement utilizing EOG procedures. If, as suggested by Moray, alterations in dwell time on particular instruments may index visual information abstraction inefficiencies, then fixation pause durations as evaluated with EOG may provide a reasonable (and inexpensive) way of obtaining such (or similar) information.

3.3 Intra-aural temperature

- Recommendation based on comments by Peter Hancock. He has been using this measure at NASA-Ames in their simulators.

- Possible measures advocated were: absolute temperature and temperature change measured from one ear; and of differential temperature between the ears.

3.3.1 Comments from principal advocate

- The measure reflects brain (perhaps hypothalamic) temperature. To the extent that metabolic "need" produces enhanced blood flow in the brain, this technique allows for the evaluation of such changes in metabolic activity.

- It is information that is inexpensively acquired, and Hancock will make his device available to the project.

3.3.2 Comments from panel participants

- We did not have available relevant resource material to evaluate the claims by Hancock. He will provide Douglas with reprints relevant to the use of this procedure in the evaluation of work load.

- There were comments about technological problems such as positioning of sensor so that it picked up tymponic membrane temperature rather than skin temperature.

- Technical problems associated with the headset worn by the pilot.

3.3.3 Potential problems

- Some are listed above.

- The panel was not strong in its endorsement of this measure be[illegible] lack of information available to us concerning [illegible] work load assessment in general and flight simulat[illegible] [illegible]ular.

4.0 Recommendations for the NEAR future: Measures recommended for consideration

4.1 Event related brain potentials (ERPs)

4.2 Eye: point-of-regard information

4.3 Voice analysis

Rationale for choice of measures

4.1.1 Event related potentials (Principal proponents - Kramer, Stern)

- The University of Illinois Cognitive Psychophysiology laboratory, in conjunction with aviation psychology (Wickens), has extensive experience with the use of this technology to evaluate aspects of work load as well as cognition. The University of Illinois investigators have focused on the use of ERPs to both primary (embedded) and secondary task stimuli, while the Washington University effort has begun to investigate the use of "irrelevant probe stimuli." All three appear to be promising, with respect to simulator applications and the embedded and probe stimuli procedures should be usable in flight environments.

- The use of embedded signals (embedded signals are part of the primary task) for the triggering of ERPs was recommended as an alternative to using secondary tasks. Other suggestions involved the use of "irrelevant" probe stimuli to elicit ERPs and the use of subject produced "responses," such as saccadic eye movements, to trigger the averaging process.

- A fall-out from recording ERPs is the EEG. Spectral analysis of this data was suggested as another window to capture levels of arousal or alertness (the other window being HR). Since alertness should not be a problem in the proposed simulation, this idea was not explored further.

- The technology for recording EEG (including ERPs) in simulators and in-flight is available and being used (ERPs - University of Illinois, Kramer; AAMRL/HEG, Wilson and EEG in flight recording, B. Sterman, centrifuge USAFSAM/Lewis). ERPs have been demonstrated to be sensitive to task demand or work load effects.

4.1.2 Problems

- Signal/noise ratio - Signal of interest is frequently degraded and lost during flight and simulation conditions.

4.1.3 Solutions

- Solutions, such as special electrodes, amplifying the signal at the source; and development of filtering procedures for signal "sensitizing," are in reasonably advanced stages of development.

4.2 Eye point-of-regard - Principal proponent - Moray.

- There is considerable literature demonstrating the utility of this technique in flight simulation (this material is available in the bibliography provided). Moray suggested, as an example, applying this technique to work load assessment, that dwell time on instruments is generally in the nature of 500 msec. If the eye returns to the same instrument frequently (i.e., average dwell time is decreased), it might suggest that the person's ability to retain information abstracted in short term memory is impaired. This might be considered an indicant of high work load.

- With respect to using the technique in flight, Moray expressed optimism about the possibility of developing the necessary instrumentation to record eye position information under such conditions. However, such instrumentation is not currently available.

4.3 Voice analysis (Principal proponent - Walrath)

- Little comment, other than that it might be a useful procedure for the future. There was some discussion about its utility to evaluate "stress," but few comments relevant to how it might be used to evaluate aspects of work load. Several articles describing laboratory validation were entered in the reference data base. Current work, sponsored by NASA/Langley will be monitored. Since pilots engage in voice communication, using voice output would be the least intrusive measure of all those sampled by us.

5.0 What physiological measures contribute to the assessment of work load.

5.1 They provide some measures that cannot otherwise be obtained (unless one can impose secondary tasks), measures that will be most useful in the evaluation of conditions involving underload of the operator. The states of concern mentioned include "arousal," "alertness," "attention," "daydreaming," "drowsiness," and "microsleep," to mention but a few.

5.2 They provide information about moment-to-moment changes in the operator, rather than average values (averaged over time). Thus, points (or narrow windows) of "momentary" overload can be identified.

5.3 They are unobtrusive and objective.

5.4 They should be used to complement other measures. As amply demonstrated during the general presentations and discussions, subjective, performance, and physiological measures are far from perfectly correlated with each other. They, therefore, provide complementary (not competitive) bits of information to those concerned with the evaluation of mental work load.

5.5 They have the potential for allowing us to discriminate between "controlled" and "automatic" information processing by the operator.

6.0 Sensitivity

The issue of sensitivity of physiological measures to graded changes in (or levels of) workload was not discussed. Our discussion focused principally on looking for differences between resting and "load" conditions. The literature, at best, has attempted to discriminate between three levels of work load (low, medium, and high).

A number of reasons (rationalizations) can be used to account for the relative lack of effort in this important area. The first is our difficulty in establishing a metric, other than an ordinal one, for defining work load levels. The second is the lack of a substantial data base relating physiological measures to work load that unequivocally demonstrates the utility of such a measure for work load assessment. The field is still young!

A third is the hope that physiological measures will provide the metric for defining work load levels with greater precision than currently possible. We should point out that our definition of "work load" may be radically different from that of the Human Factors Engineer. It is our suspicion that their preferred definition deals principally with the load imposed on an operator by a given configuration of hardware. Our preferred definition is in terms of the impact of the imposed load on an operator. Thus, the point in the information processing chain that is sampled by those who attempt to define work load on the basis of what is imposed on the operator is different from those who focus on the impact of the imposed load on aspects of operator performance. In our case, the performance measure is the output from one or more physiological systems.

7.0 Warning comments

7.1 At the current stage of development, those enlisted to collect physiological data must be trained: to discriminate signal from noise; in the proper application of electrodes; and in appropriate signal conditioning procedures (amplification, filtering, etc.).

7.2 Environments in which bio-electric data is to be collected have to be "sanitized," i.e., sources of electrical artifact have to be shielded, equipment properly grounded, etc.. This may require the services of a biomedical engineer.

7.3 Data reduction is, at best, a semi-automatic process.

7.4 Data collection and reduction is a relatively costly procedure, when compared to paper and pencil tests or recording the outputs from "manipulanda" or in-flight equipment.

E X P E R I M E N T A L   D E S I G N

SESSION ONE: SWAT in-flight NASA/TLX post-flight

ORDER NUMBER: 1	2	3	4
SFO - SCK (LO)	SMF - SFO (HI)	SMF - SFO (LO)	SFO - SCK (HI)
SMF - SFO (HI)	SFO - SCK (LO)	SFO - SCK (HI)	SMF - SFO (LO)

SESSION TWO:

SMF - SFO (HI)	SFO - SCK (LO)	SFO - SCK (HI)	SMF - SFO (LO)
SFO - SCK (LO)	SMF - SFO (HI)	SMF - SFO (LO)	SFO - SCK (HI)
Ss 1,9,17	2,10,18	3,11	4,12

SESSION ONE: NASA/TLX in-flight SWAT post-flight

ORDER NUMBER: 1	2	3	4
SFO - SCK (LO)	SMF - SFO (HI)	SMF - SFO (LO)	SFO - SCK (HI)
SMF - SFO (HI)	SFO - SCK (LO)	SFO - SCK (HI)	SMF - SFO (LO)

SESSION TWO:

SMF - SFO (HI)	SFO - SCK (LO)	SFO - SCK (HI)	SMF - SFO (LO)
SFO - SCK (LO)	SMF - SFO (HI)	SMF - SFO (LO)	SFO - SCK (HI)
Ss 5,13	6,14	7,15,19	8,16,20

M E A S U R E M E N T   W I N D O W S

FLIGHT PHASE	OPEN MEASUREMENT WINDOW	CLOSE MEASUREMENT WINDOW
• Takeoff:	EPR .GT. 1.50	FLAPS 5
• Climb:	FLAPS UP	1:00 (One minute later)
• Cruise 1:	10,000 ft	2:00 (Two minutes later)
• Cruise 2:	3 minutes after 10,000 ft	2:30 (Two and a half minutes later)
• Descent:	Throttles to Idle	5,500 ft
• Approach:	Localizer activiation	Outer Marker
• Touchdown:	Middle marker	1:30 (One and a half minutes later)

5 2.5 *SEES ALTITUDE ALERT LIGHT ILLUMINATE.
5 *HEARS ENGINE SURGE NOISES.
4 9(8a) *HEARS F/E, "NO. 3 ENGINE IS SURGING."
5 9(2) 9(5) *NOTES #3 EPR, N1 AND N2 FLUCTUATING.
5 9(2) 9(5) *NOTES #3 EGT RISING.
4,6 9(3) 9(5) 9(8a) *CAPT TELLS F/O, "BRING #3 TO IDLE."
1,5 9(1) 9(2) *CAPT TRIMS RUDDER AND STABILIZER. NOTES F/O *
5 9(5) BRINGING #3 THROTTLE TO IDLE.
4 9(8a) *HEARS F/E, "#3 IS STILL SURGING. WE APPEAR TO BE GETTING COMPRESSOR STALLS."
4 9(2) 9(5) *CAPT CALLS, "ROGER, LET'S SHUT IT DOWN. ENGINE SHUTDOWN CHECKLIST."
1,6 9(1) 9(2) 9(3) *CAPT ADDS POWER, TRIMS RUDDER AND STABILIZER.
1 / 9(1) * CAPTAIN ADDS LEFT RUDDER

00:09:20 ENGINE FAILURE/SHUTDOWN CHECKLIST.

4,5 7 9(8a) *HEARS F/E, "ESSENTIAL POWER OPERATING GENERATOR."
4,5 7 9(8a) *HEARS F/E, "THRUST LEVER -- CLOSED."
5 7 9(3) *CAPT ~~CLOSES~~ NO. 3 THRUST LEVER. CHECKS (WAS CLOSED HERE
4,5 7 9(8a) *HEARS F/E, "START LEVER -- CUTOFF."
5 7 9(1) 9(3) *CAPT PLACES START LEVER TO CUTOFF.
4,5 7 9(8a) *F/E SAYS, "CARGO OUTFLOW VALVE -- CLOSED."
4,5 7 9(8a) *F/E SAYS, "GENERATOR BREAKER LIGHT -- ON."
4,5 7 9(8a) *F/E SAYS, "ELECTRICAL LOADS -- MONITOR."
4,5 7 9(8a) *F/E SAYS, "ENGINE BLEED AIR SWITCH -- CLOSED."
4,5 7 9(8a) *F/E SAYS, "WING AND ENGINE ANTI-ICE SWITCHES -- CLOSED."
4,5 7 9(8a) *HEARS F/O RESPOND, "CLOSED."
4 9(3) 9(8a) *HEARS F/E, "ENGINE SHUTDOWN CHECKLIST COMPLETE."

00:10:00 LEVEL OFF AT 11,000 FT AND CROSS RESAS INTERSECTION.

5 2,5 *SEES ALTITUDE ALERT LIGHT EXTINGUISH.
4,6 9(3) 7 9(8a) *TELLS F/E TO SET CRUISE THRUST FOR ENGINE INOP CRUISE.
4 9(8a) *HEARS F/E, "POWER SET."
1,6 9(1) 9(3) 7 *CAPT TRIMS RUDDER AND STABILIZER FOR CRUISE.
4 9(3) 9(8a) *HEARS F/E, "ONE GENERATOR INOP CHECKLIST IS NOW COMPLETE."
4 9(3) 9(8a) *CAPT ACKNOWLEDGES.
1,3,6 1,2,8b *LEVELS AIRPLANE AT 11,000 FEET USING ADI, ALTIMETER AND VERTICAL SPEED INDICATOR.
1,5,6 1,3 *PLACES FD ~~AND AUTOPILOT~~ ALTITUDE HOLD SWITCH ON.
5,6 1,2,5 *ACCELERATES TO 288 KNOTS. (2 ENGINE CRUISE)
4,6 3,8a *CALLS, "LANDING LIGHTS OFF."
5,6 1,2,3 *SETS AIRSPEED CURSOR TO 288 KNOTS.
5 2,5 *OBSERVES APPROACHING DESIRED AIRSPEED.
5,6 1,2,3 *SETS THRUST TO MAINTAIN DESIRED AIRSPEED.
3,5 2,8b *OBSERVES DISTANCES TO MODESTO ON DME INDICATOR.
3,5 2,8b *OBSERVES 088 DEG COURSE INTERCEPT TO MODESTO VORTAC.
1,3,5,6 1,2,3,8b *INITIATES RIGHT TURN ON HSI AS COURSE DEVIATION INDICATOR APPROACHES CENTER.

SUBJECT [illegible]

WINDOW [illegible]

FUNCTION

1. FLIGHT PATH 1
2. COLL. AVOID. 2
3. NAVIGATION 0
4. COMMUNICATION 15
5. OP & MONITOR 3
6. COMMAND DEC. 2

FACTOR

1. CONTROLS 0
2. DISPLAYS 0
3. PROCEDURE 1
4. XXXXXXXXXXXXXXXXXXXXXX
5. MONITOR 0

8A. COMMUNIC. 2

8B. NAVIGATE 0

7. AUTO 13

NON-NORMALS

9(1) NN CONTROL 3

9(2) NN DISPLAY 0

9(3) NN PROCED. 6

9(5) NN MONIT. 0

9(8A) NN COMM. 13

9(8B) NN NAV. 0

P 11

M 30

T A S K   L O A D I N G

*** PHYSICAL / MENTAL ***

	LOW SFO-SCK	LOW SMF-SFO	HIGH SFO-SCK	HIGH SMF-SFO
T/O	**75** / 124	**71** / 117	**75** / 124	**71** / 117
CLIMB	**39** / 75	**38** / 76	**40** / 82	**39** / 95
			AUTOPILOT INOP	
CRUISE1	**13** / 46	**10** / 37	**34** / 128	**31** / 121
			NO# 3 ENGINE COMPRESSOR STALL	
CRUISE2	**1** / 6	**1** / 5	**11** / 50	**11** / 49
			B SYSTEM HYDRAULICS FAILURE	
DESCENT	**47** / 137	**45** / 141	**46** / 140	**53** / 164
APPROACH	**63** / 185	**66** / 199	**61** / 177	**64** / 190
LAND	**28** / 62	**28** / 62	**26** / 59	**26** / 59
TOTALS	**266** / 635	**259** / 637	**293** / 760	**295** / 795
	901	896	1053	1090

EXPERIMENTAL VARIABLES

4 FACTORS: 2 X 2 X 4 X 7 DESIGN

FACTOR ONE: WORKLOAD (2 levels) LOW and HIGH

FACTOR TWO: ROUTE (2 levels) SFO - SCK & SMF - SFO

FACTOR THREE: FOUR POSSIBLE ORDERS A, B, C, & D

FACTOR FOUR: MEASUREMENT EPOCH **(TASK LOADINGS)** (7 levels)

1) Takeoff
2) Climb
3) Cruise 1
4) Cruise 2
5) Descent
6) Approach
7) Landing

DATA ANALYSES

SENSITIVITY ANALYSES

CONTENT VALIDITY

2 X 7 ANOVA Workload by Measurement Epoch **(TASK LOADINGS)**

- o Various Dependent Variables, one at a time
- o MANOVA approach presupposes a battery of tests

Then a series of planned comparisons:

Considerations:
- o Maintain comparison-wise Type I error rate
- o Select a test which is ROBUST to departures from equal variances and unequal sample sizes

D A T A   A N A L Y S E S

CONTENT VALIDITY

o **SENSITIVITY TO DIFFERENT PHASES OF FLIGHT**

Within a Flight Segment (i.e., SFO - SCK) compare different Measurement Epochs **(TASK LOADINGS)**

SFO - SCK Takeoff (LOW)
to
SFO - SCK Cruise 1 (LOW)

* * * * * * * * * * * *

QUESTION: "Can the workload measure discriminate between different phases of the same flight segment?"

DATA ANALYSES

CONTENT VALIDITY

o **SENSITIVITY TO DIFFERENT FLIGHT SEGMENTS (HIGH AND LOW TASK LOADINGS)**

Between Flight Segments (i.e., LOW & HIGH) compare different Levels of Workload

SFO - SCK Descent (LOW)
to
SFO - SCK Descent (HIGH)

* * * * * * * * * * * *

QUESTION: "Can the workload measure discriminate between the same phase of the different flight segments?"

D A T A   A N A L Y S E S

ALTERNATE FORMS / TEST-RETEST RELIABILITY

o SESSION ONE TO SESSION TWO

Collapsed across Flight Segments, compare different Measurement Epochs **(TASK LOADINGS)**

SFO - SCK Descent (LOW)
to
SMF - SFO Descent (LOW)

* * * * * * * * * * * * *

QUESTION: "Is the measure stable? Will the same **relative differences** in workload be found with repeated testing?"

DATA ANALYSES

CONSTRUCT VALIDITY

- CORRELATION COEFFICIENTS COMPARING VARIOUS WORKLOAD MEASURES

  Collapsed across Measurement Epochs **(TASK LOADINGS)**, compare the same Flight Segments

SFO - SCK (HIGH)
Physiological Measure #1
to
SFO - SCK (HIGH)
Subjective Measure #1

* * * * * * * * * * * *

QUESTION: "Are different workload measures sensitive to the same variations in TASK DEMANDS?"

Simulation Facility

- NASA - Ames (MVSRF)
- Boeing 727 motion-base simulator
- High level of fidelity
- ATC simulation

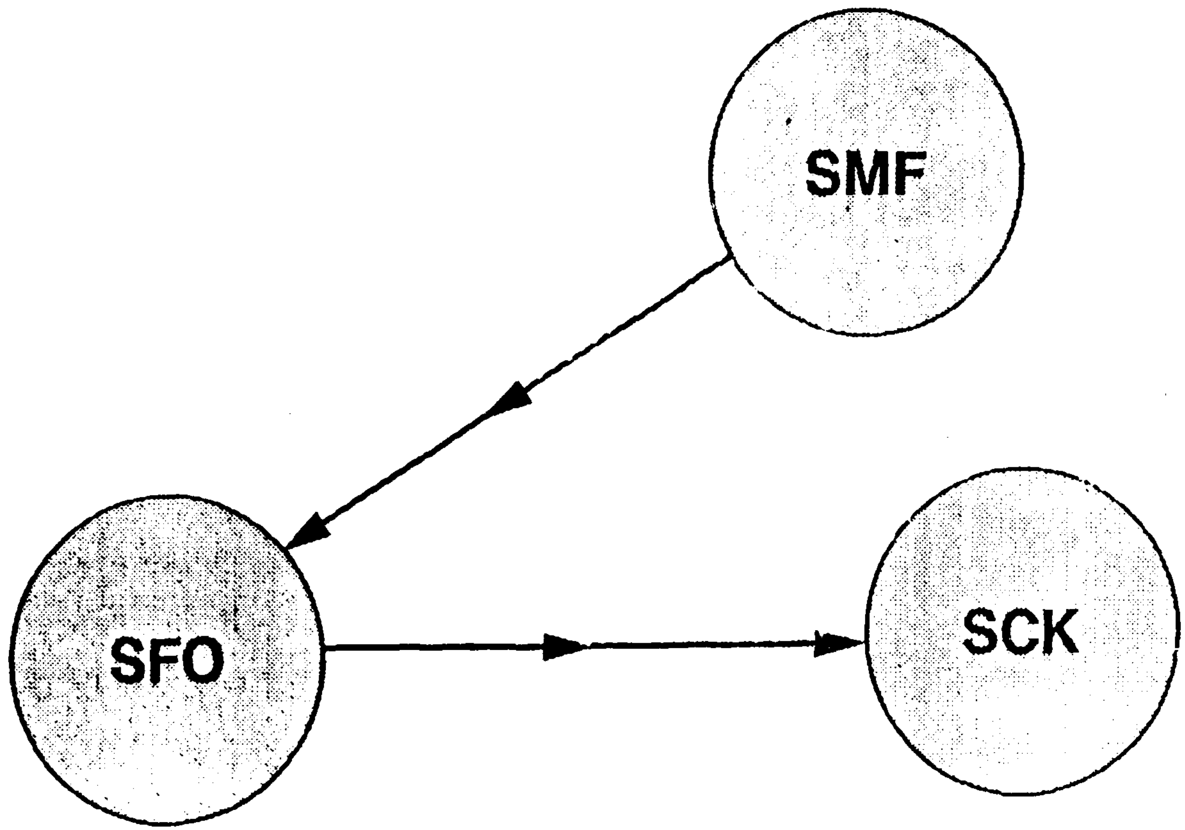
A1167.01

Subjects

- FAR qualified and current 727 airline pilots
- Captain (data collection)
- Confederates
- Representative of population

A116/.02

Simulation Scenarios



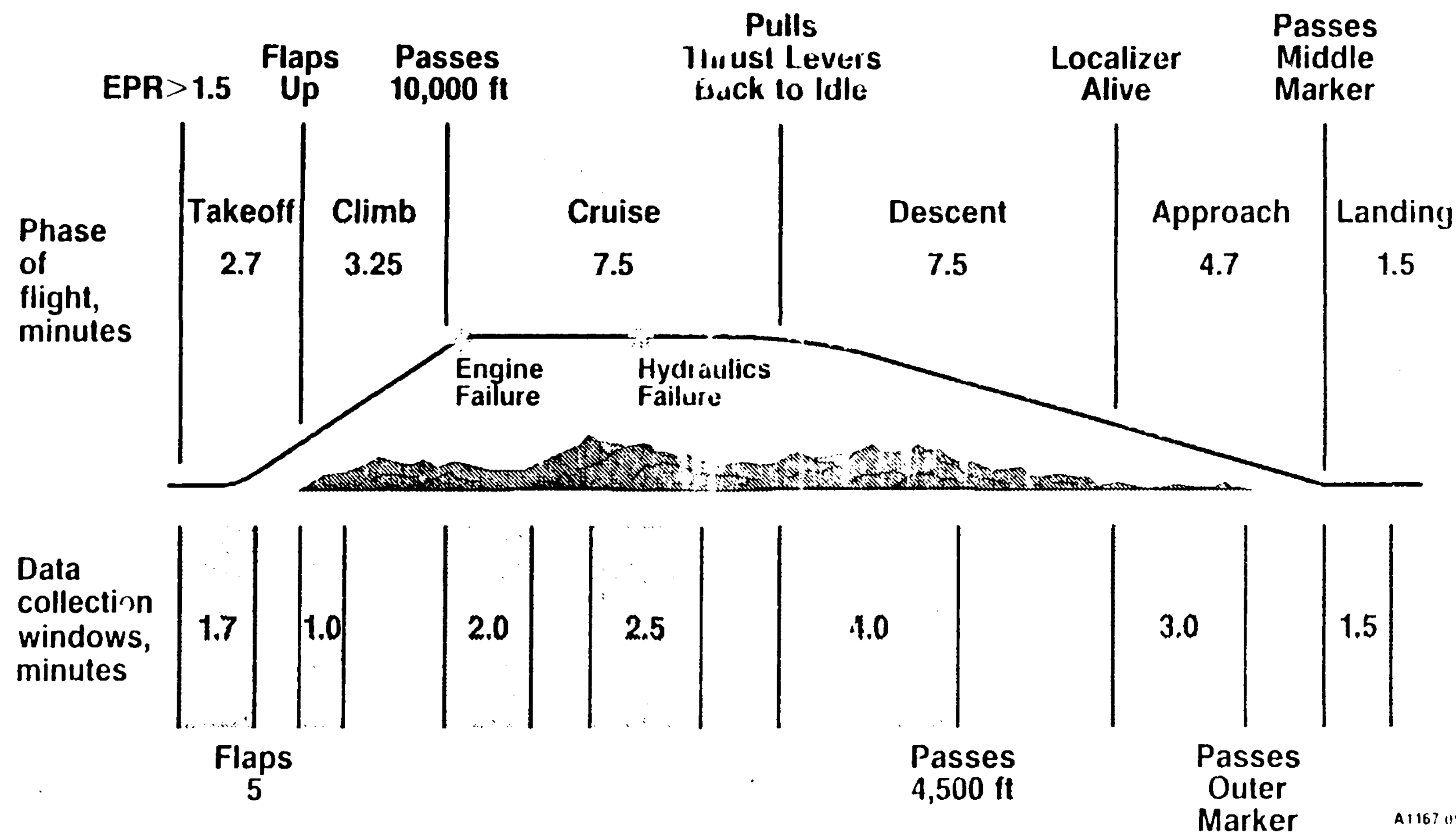
A116/63

Two Workload Levels

Conditions	Level	
	High	Low
Weather	Ceiling, 500 ft; visibility, 1 mi	Clear
Wind	12 kn at takeoff and landing	5 kn at takeoff and landing
Autopilot	Inoperative	Operating
Turbulence	Significant	Minimal
Nonnormals	• Number three engine stall • Hydraulic System B failure • Distractors (i.e., autopressure failure, window overheat)	None

A1167 04

Phases of Flight and Data Collection



Operationally Relevant Types of Workload
FAR 25.1523, Appendix D

- FAA addresses
  - 6 workload functions
  - 10 workload factors
- Map FAR-25 function and factor descriptions into scenarios
- Divide scenarios into high and low workload levels based on objective task demands

A1167.06

Basic Workload Functions

FAR-25

1. Flightpath control
2. Collision avoidance
3. Navigation
4. Communications
5. Operations and monitoring of aircraft engines and systems
6. Command decisions

A116/07

Workload Factors

FAR-25

1\. Controls
2\. Displays
3\. Procedures
4\. Mental and physical effort
5\. Monitoring
8\. Communication and navigation

9\. Nonnormals
6\. Crew member out of area
7\. Automation breakdown
10\. Incapacitated crew member

A1167 08

Function and Factor Mapping

Example

00:02:15 Function	Factor	Gear Retract - Start Initial Climb
4,5,6	3,8A	• Calls, "gear up"
4	8A	• Hears F/O, "gear up"
5	5	• Sees F/O compliance (peripheral vision)
1,5	2,3,5	• Looks to see if airspeed stabilized at approximately V2 + 10
1,5,6	1,2,3	• Adjusts pitch attitude to maintain V2 + 10 if necessary
1,5,6	1,2,3	• Sets FD pitch knob to proper pitch attitude
00:02:35		Cleared Direct Sacramento Vortac

A1167 09

Workload Factors

1a. Accessibility of controls
1b. Ease of controls
1c. Simplicity of controls

2a. Accessibility of displays
2b. Conspicuity of displays
2c. Interpretability of displays

3a. Urgency of procedures
3b. Ease of procedures
3c. Simplicity of procedures

A116/.1

Workload Factors

4a. Degree of concentrated mental effort
4b. Duration of concentrated mental effort
4c. Degree of concentrated physical effort
4d. Duration of concentrated physical effort

5. Amount of required monitoring

8a. Amount of communication
8b. Amount of navigation

9. Nonnormals
6. Crew member out of area
7. Automation breakdown
10. Incapacitated crew

A1167 22

Factors

1. Controls
   - Any manipulation of any aircraft control (e.g., switches)
   - If Factor 1 is assigned, Factor 5 cannot be used

2. Displays
   - Any visual confirmation or visual reference to an indicator showing the state of the aircraft (e.g., CRTs, dials, etc.)

3. Procedures
   - Any standard (normal) action for normal operation of the aircraft

5. Monitoring
   - The extent of required monitoring for normal aircraft operation
   - Factor 5 excludes the use of Factors 1, 8a, and 8b and Functions 3 and 4

A1167 17

Functions

1. Flightpath control
   - Anytime crew senses movement of the airplane (e.g., pilot senses airplane start to roll)
   - If the crew manipulates anything to cause any up-down or left-right motion

2. Collision avoidance
   - If navigating, and crew looks out window for anything
   - If ATC directs an action

3. Navigation
   - Any altitude or heading information
   - If Function 3 is assigned, Factor 8b must be used and Factor 5 cannot be used

A1167.19

Factors

8a. Communications
- Any Function 4 fires Factor 8a
- Factor 8a excludes the use of Factor 5

8b. Navigation
- Any Function 3 fires Factor 8b
- Factor 8b excludes the use of Factor 5

9. Nonnormals
- Broken down into Factors 9-1, 9-2, 9-3, 9-5, 9-8a, and 9-8b

6. Crew member out of area

7. Nonnormal conditions that require manual as opposed to automatic control of aircraft systems

10. Crew member incapacitated

A1167 13

Functions

4. Communications
   - Any verbal, incoming or outgoing speech, occurs
   - If Function 4 is assigned, Factor 8a must be used and Factor 5 cannot be used

5. Operations and monitoring of aircraft engines and systems
   - Visual confirmation of information
   - Manipulation of the aircraft (nonflightpath)

6. Command decisions
   - Anytime the pilot makes an action to manipulate the aircraft that requires a decision

A1167.20

Dependent Measures

- Subjective
- Physiological
- Performance
- Task timeline analysis

A1167 11

Subjective Measures

- Inflight (direct measurement)
- Postflight (videotape)

- Example measures
  - NASA TLX
  - SWAT
  - Modified Cooper-Harper

A1167.12

Physiological Measures

- Heart
  - Mean HR
  - SD
  - Spectrum
- Eye blinks
  - Number
  - Mean duration
  - SD duration
- Eye movements
  - Number
  - Fixation
  - Mean duration
  - SD duration
  - EOG rate
- EEG

A1167 1

Performance Measures

- Reversals
  - Stick
  - Aileron
  - Rudder
  - Throttle

- Approach and landing
  - Glideslope error
  - Flight director error
  - Glideslope and localizer variability
  - Altitude at outer, middle, and inner markers

A1167 14

Task Timeline Analysis (TLA)

- Selected segments
- Proven TLA method
- Identifies high and low task-demand levels
- Validity of workload measures against proven tool

A116/15

Performance Measures

- Reversals
  - Stick
  - Aileron
  - Rudder
  - Throttle

- Approach and landing
  - Glideslope error
  - Flight director error
  - Glideslope and localizer variability
  - Altitude at outer, middle, and inner markers

A1167 14

Task Timeline Analysis (TLA)

- Selected segments
- Proven TLA method
- Identifies high and low task-demand levels
- Validity of workload measures against proven tool

A1167 15

Concern Areas

- Aircraft certification precludes implementation of many secondary tasks for performance measures
- Physiological and performance measures → inflight
  Subjective measures → postflight
- Data collection windows are of different lengths

A1167 1b

ASSESSMENT OF CREW WORKLOAD WORKSHOP SURVEY

NAME______________________

MEASURE PREFERENCES

Because your views may be different from those already presented by the speakers, or, from those presented from your working group, we are interested in what you believe to be the best workload measures. If your views are different from those already presented, please indicate which measures you believe are best, in order, and provide a rationale for your view.

Measure Type: Subjective Physiological Performance

Measure	Rationale
1.________________	
2.________________	

SIMULATION RECOMMENDATIONS

How would you improve the design of the part-task Simulation?

SUGGESTED MEASURES

Questionnaire Summary

GROUP	MEASURE	COUNT
Physiological	Voice Measures	3
Physiological	Auditory Canal Temperature	2
Physiological	ERP	2
Physiological	EEG Spectra	1
Physiological	Eye Movement	1
Subjective	Modified Cooper-Harper	2
Subjective	Airbus 2-8 Scale	1
Subjective	Bedford Scale	1
Subjective	Boeing Comparative	1
Subjective	In Flight, High Level Subjective	1
Perform. Sec.	Time Estimation	2
Perform. Sec.	Embedded Communication	1
Perform. Sec.	Unobtrusive Embedded Secondary Task	1
Perform. Prim.	Primary Task (mission effectiveness)	1
Other	Performance/Physiological	1
Other	SSER	1
Other	Subjective/Physiological	1
Other	Timeline Analysis	1

Questionnaire Summary

SIMULATION SUGGESTIONS

SUGGESTED ADDITIONS:

- o Increase the amount of mechanical failures (switch failure to hold, CB fails).
- o Pre-correlate workload with timeline analysis results to confirm validity.
- o Use glideslope as a primary task measure for AZ, ALT deviation, power required (weighted differently across the scenario).
- o Include a <u>very</u> high workload condition to be sure the measures are measuring anything.
- o Use filtering or other 0.1Hz analysis.
- o Use NDB approach or LDA approach to load the captain.
- o Address crew members as captain, first officer, and second officer.
- o Manipulate communications workload.
- o Perform an apriori test for level factor as a 2X2 (3 treatments with 1 within subjects factor) MANOVA. If there is no effect, it can be eliminated as a factor for later analysis.
- o Use synthetic tasks early on, then embedded tasks later in full system flight tests.
- o Design a new simulator with built-in data systems and flexible, quick-change capabilities.
- o Design criterion tasks with built-in measurement schemes.
- o Measure co-pilot workload.
- o Test measures in an actual 727 flight at some time.
- o Include more ECG data in 1.0, 1.7, and 1.5 minute segments for FFT analysis.
- o Make windows longer to allow heart rate variability to stabilize.

SUGGESTED ELIMINATIONS:

- o Eliminate throttle reversal (measures technique, not workload).
- o "Improve the scenario"
- o Eliminate 2 workload states.
- o Do not use 727 for testing. Old model, will probably never be certified again. Can you be certain of validity for a highly automatic plane?
- o Eliminate compressor stall at 10,000 feet (unrealistic, should be at a higher altitude).
- o Do not assume equal workload at the 3 airports selected.

SUGGESTED CONSIDERATIONS:

- o Assymetric performance transfer
- o Range effects X stress differential interaction
- o New generation A/C
- o Full crew interaction
- o Sensitivity of measures
- o Between subjects and between test run variability
- o Anticipation between high and low levels
- o What part task simulation?
- o Mechanical malfunctions load second officer more than captain.
- o Pitfalls

SUGGESTED REFERENCES

IN REFERENCE TO MEASURES:

- o Speyer's handouts [Speyer]
- o Otis Elevator (c. 1965-1970) [Parks]

IN REFERENCE TO SCENARIO:

- o Poulton and Freeman (1973) [Hancock]
- o WPAFB, Human Factors Branch, Crew Station Designs Division c. 1979 (call Richard Gesselhart (513) 255-4109) [Metcher]
- o O'Donnell and Eggemeier; Gopher and Donchin (current) [Derrick]

END
FILMED
MARCH, 1988
DTIC